Как написать свой индекс в Tarantool

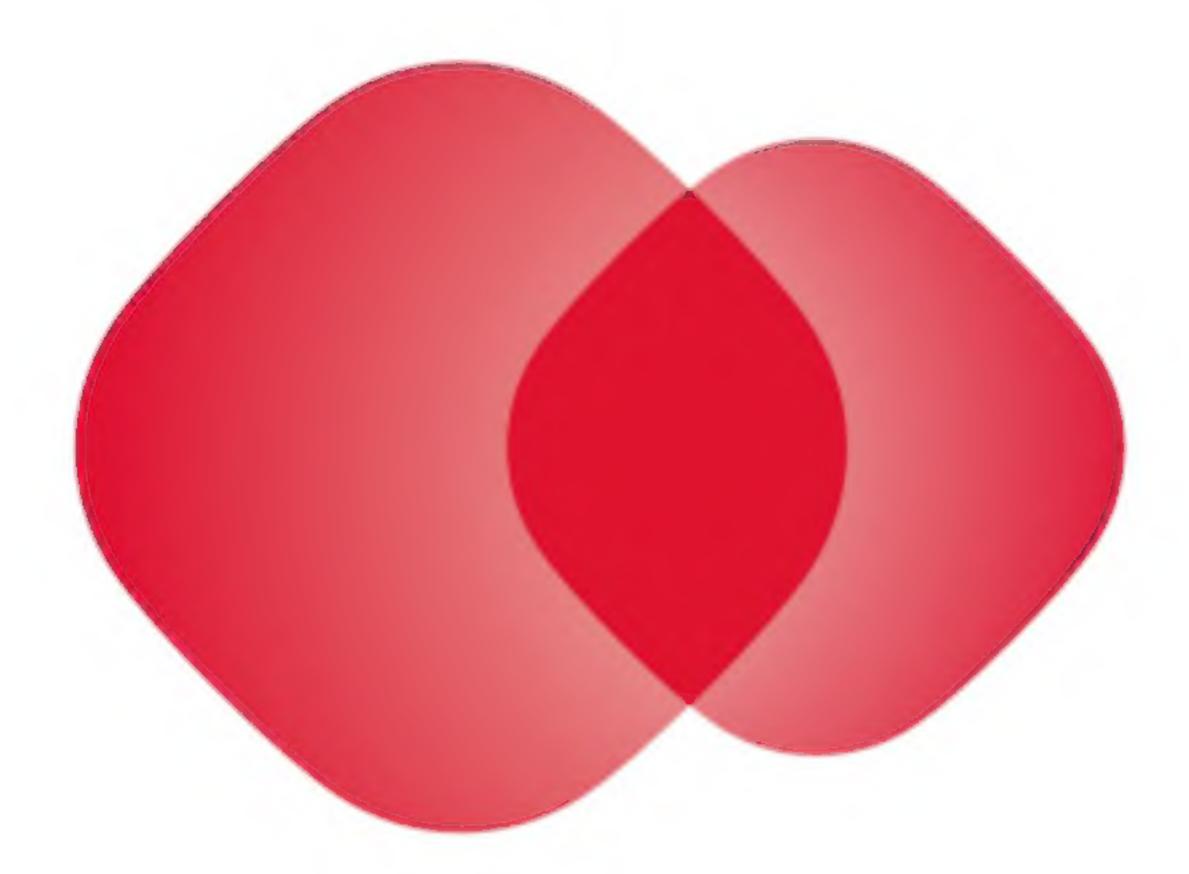
Бабин Олег



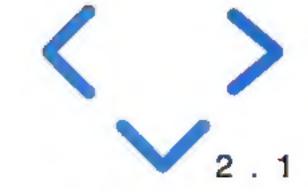


Tarantool

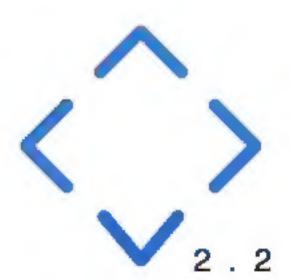
- Сервер приложений
- База данных
 - memtx хранение в памяти
 - vinyl хранение на диске



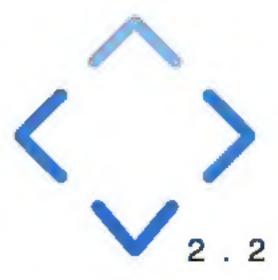




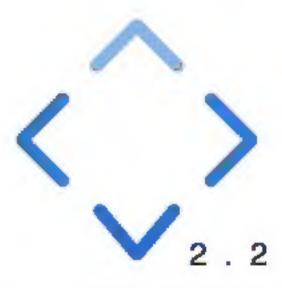




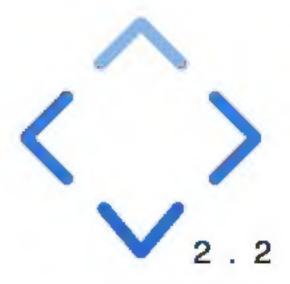




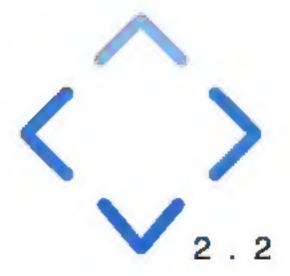








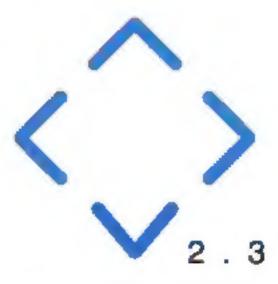




Выборка данных

```
tarantool> box.space.my space:select()
3 - - [1, 'hello']
 - [2, 'world']
 tarantool> box.space.my space:select({1},
                 {iterator = box.index.GT})
9 - - [2, 'world']
```

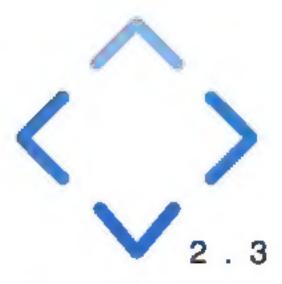




Выборка данных

```
1 tarantool> box.space.my space:select()
3 - - [1, 'hello']
4 - [2, 'world']
6 tarantool> box.space.my space:select({1},
                 {iterator = box.index.GT})
9 - - [2, 'world']
```

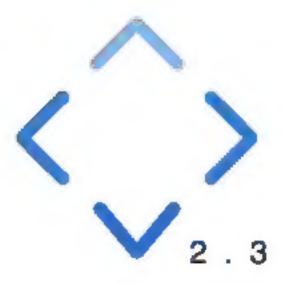




Выборка данных

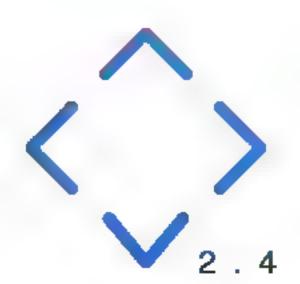
```
tarantool> box.space.my space:select()
3 - - [1, 'hello']
 - [2, 'world']
 tarantool> box.space.my space:select({1},
                 {iterator = box.index.GT})
9 - - [2, 'world']
```





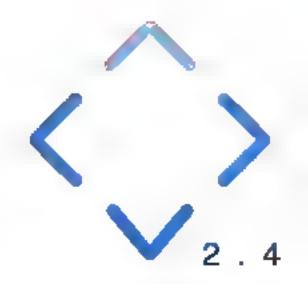
Вторичные индексы





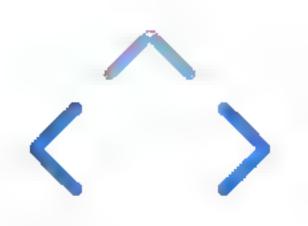
Вторичные индексы





Вторичные индексы





Многомерные данные





1 из 6 фата ...

Смартфон Xiaomi Redmi 9 4/64GB (NFC)

160 от зь вов

долгое время работы, объем памяти

Android 10 поддержка двух SIM-карт

экран 6.53°, разрешение 2340х1080 4 камеры: широкоугольная (13 МП), сверхширокоугольная,

макро, датчик глубины процессор MediaTek Helio G80

6 606 человек купили этот товар



1 из 6 фото . . .

Смартфон Apple iPhone 11 128GB

4.7 319 отзывов

🐞 качество фотографий, мощный процессор

iOS 13

поддержка двух SIM-карт (nano SIM+eSIM) **экран 6.1", разрешение 1792x828** двойная камера: 12 МП, 12 МП процессор Apple A13 Bionic

8 922 человека купили этот товар



1 из 6 фото . .

Смартфон Samsung Galaxy A71 6/128GB

273 отзыва

🕩 экран, объем памяти

Android 10 поддержка двух SIM-карт экран 6.7°, разрешение 2400x1080 4 камеры: 64 МП, 12 МП, 5 МП, 5 МП процессор Qualcomm Snapdragon 730

6 011 человек купили этот товар



Смартфон Xiaomi Redmi 9A

80 отзывов

11 599 P

от 437 ₽/мес

75 предложений от 11 528 ₽

Четкий экран, яркости мне хватает Наклеенная защитная пленка "с завода".... Андрей В.

55 390 P

327 предложений от 55 390 ₽

Уникальный аппарат, который уверенно смотрит в будущее, да и сам - уже в... Ольга К.

22 650 P

140 предложений от 22 450 ₽

Большой, красочный экран Время автономной работы Быстрая зарядка...

Максим Т.



92 плепложения от 7 339 8

Производитель

Apple BQ HONOR HUAWEI Nokia OnePlus OPPO Philips realme Samsung Xiaomi ZTE

Показать все

√ В продаже

Тип

кнопочный телефон смартфон

Платформа

Android iOS

Игровой

Диагональ экрана (точно), "

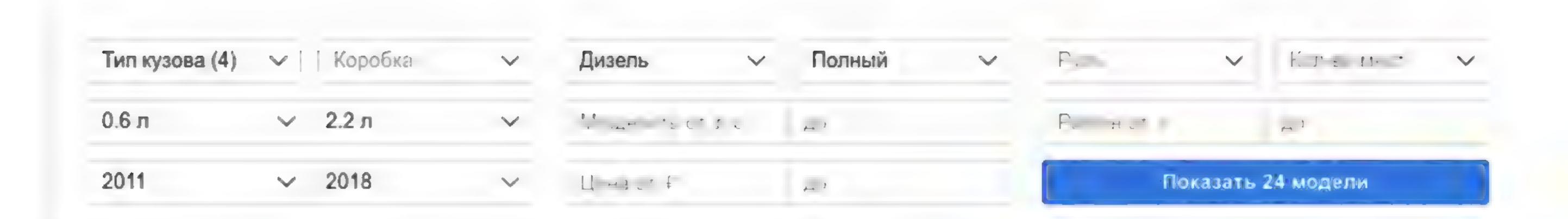
OT 0 66 до 76

Диагональ экрана

до 3.4° 35"-49" 5.0"-5.4" 5 5*-5 9* 6.0"-6.4" 65° и больше







Популярные

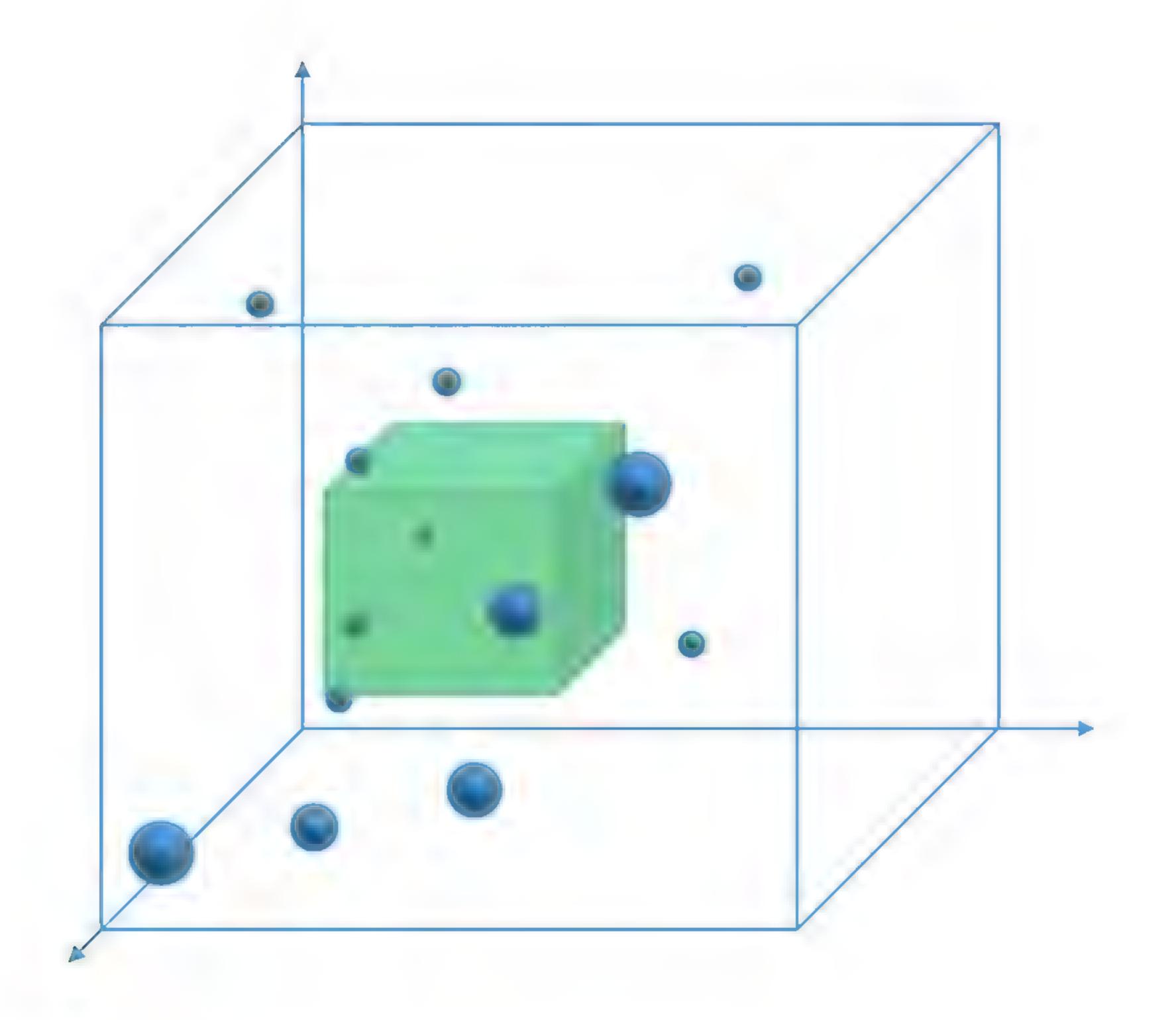




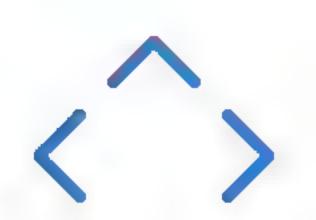




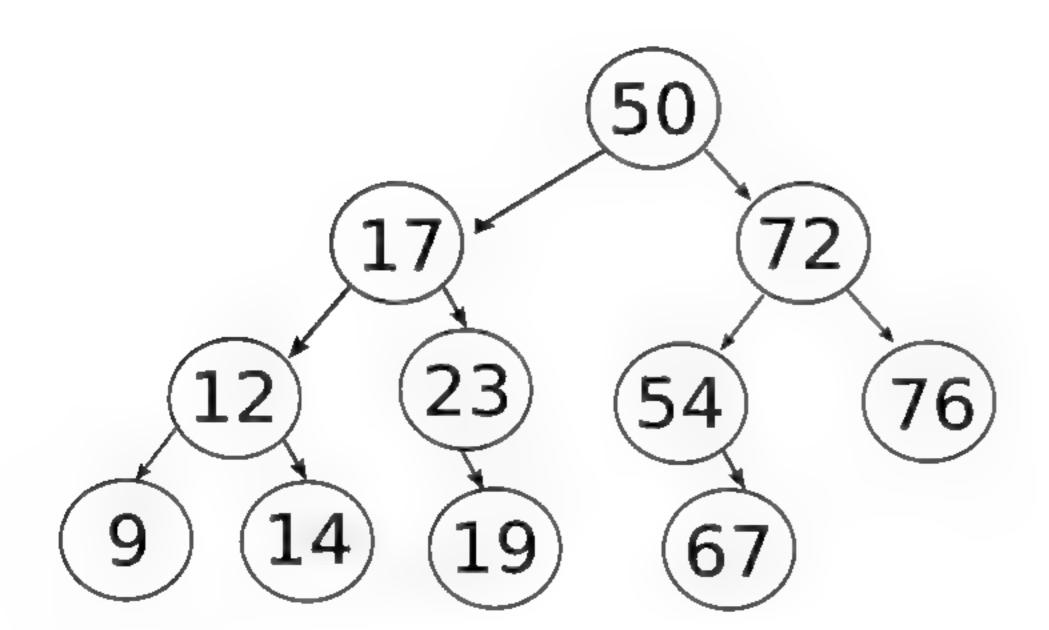
Многокритериальный поиск

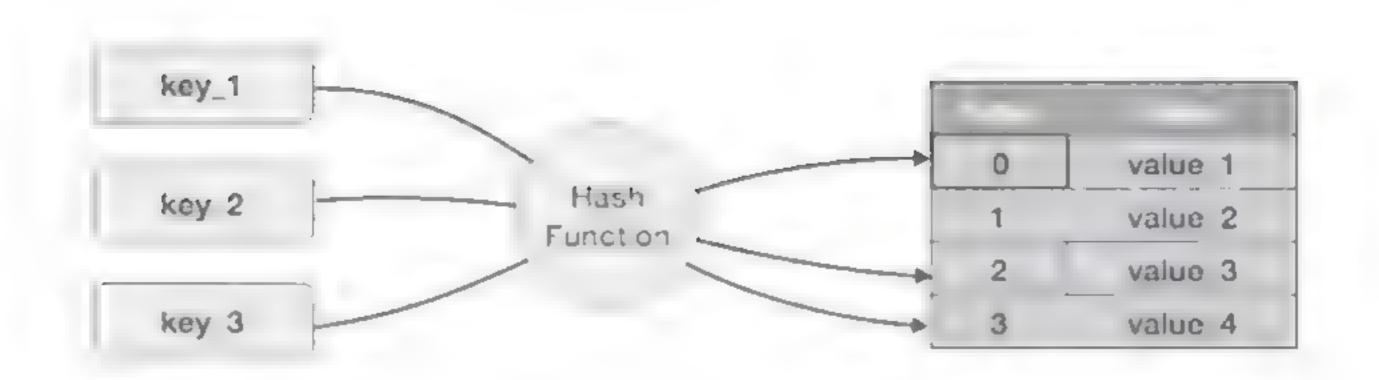






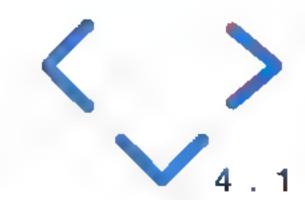
Индексы



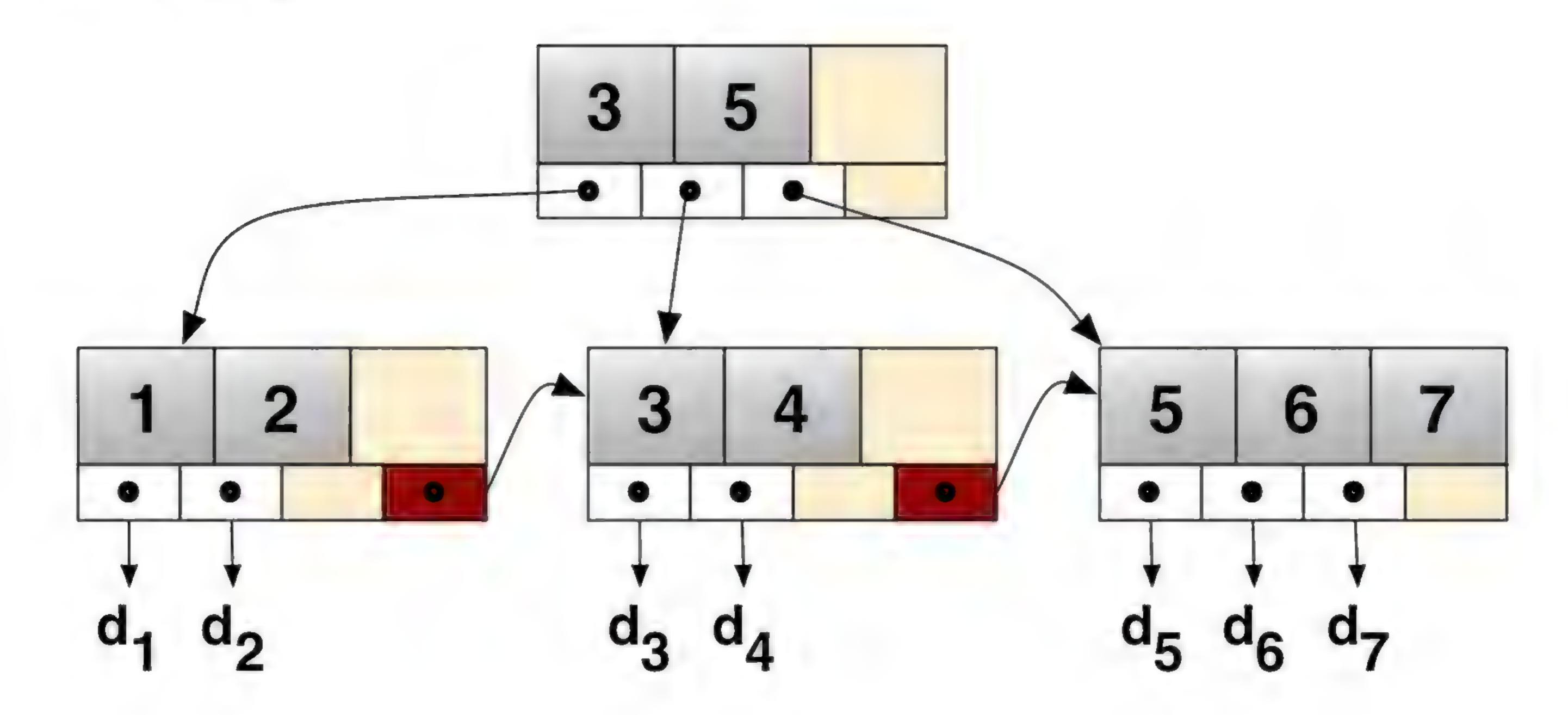




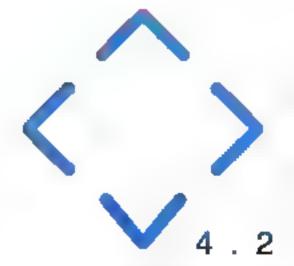


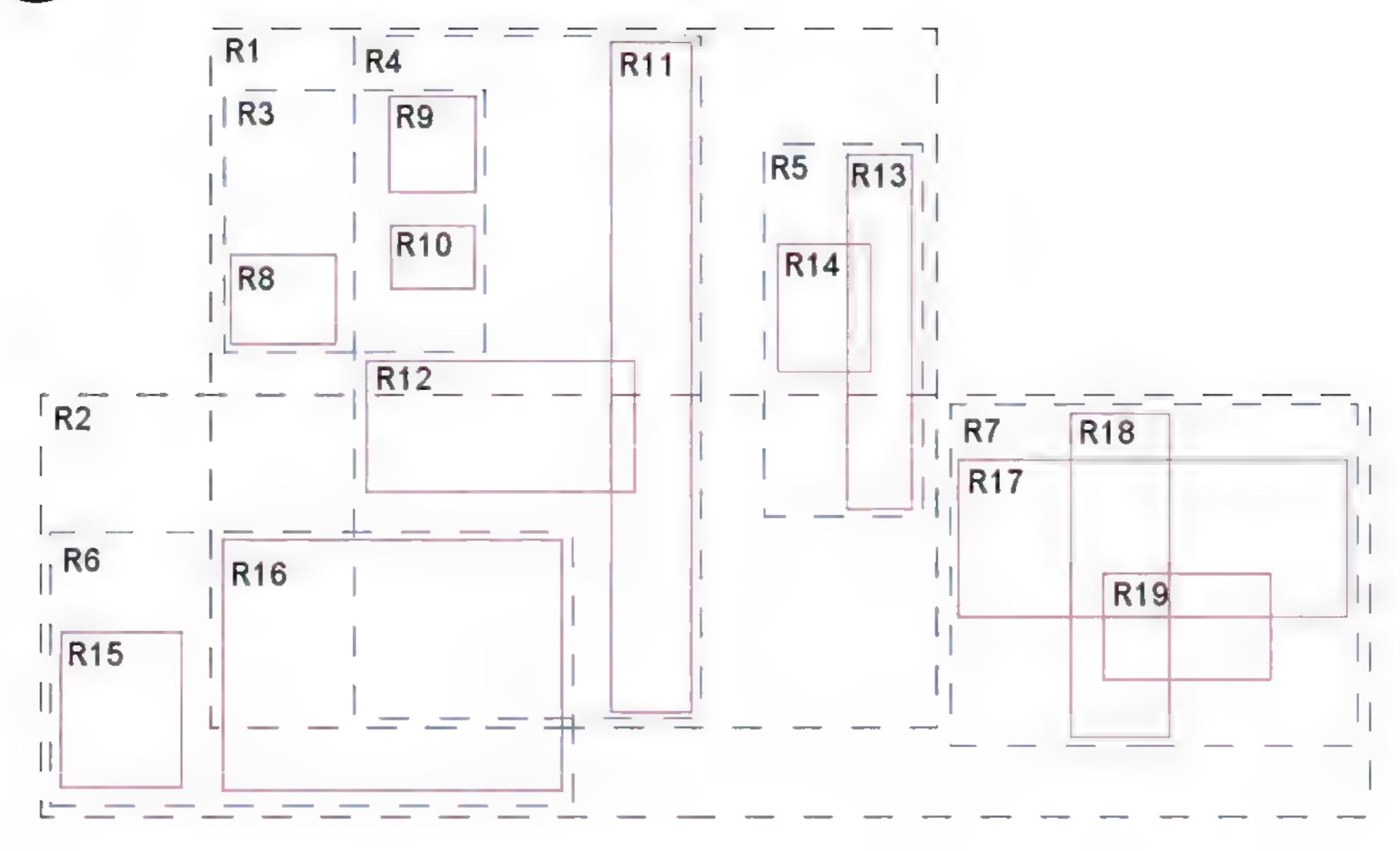


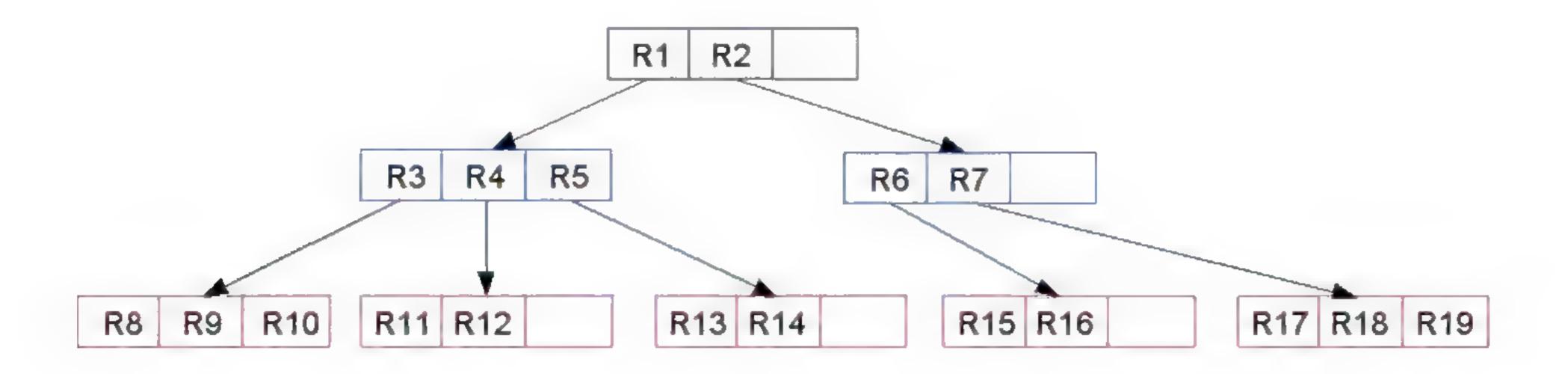
B-Tree



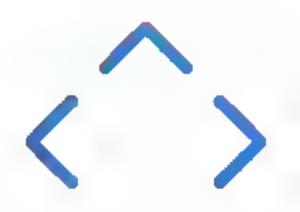






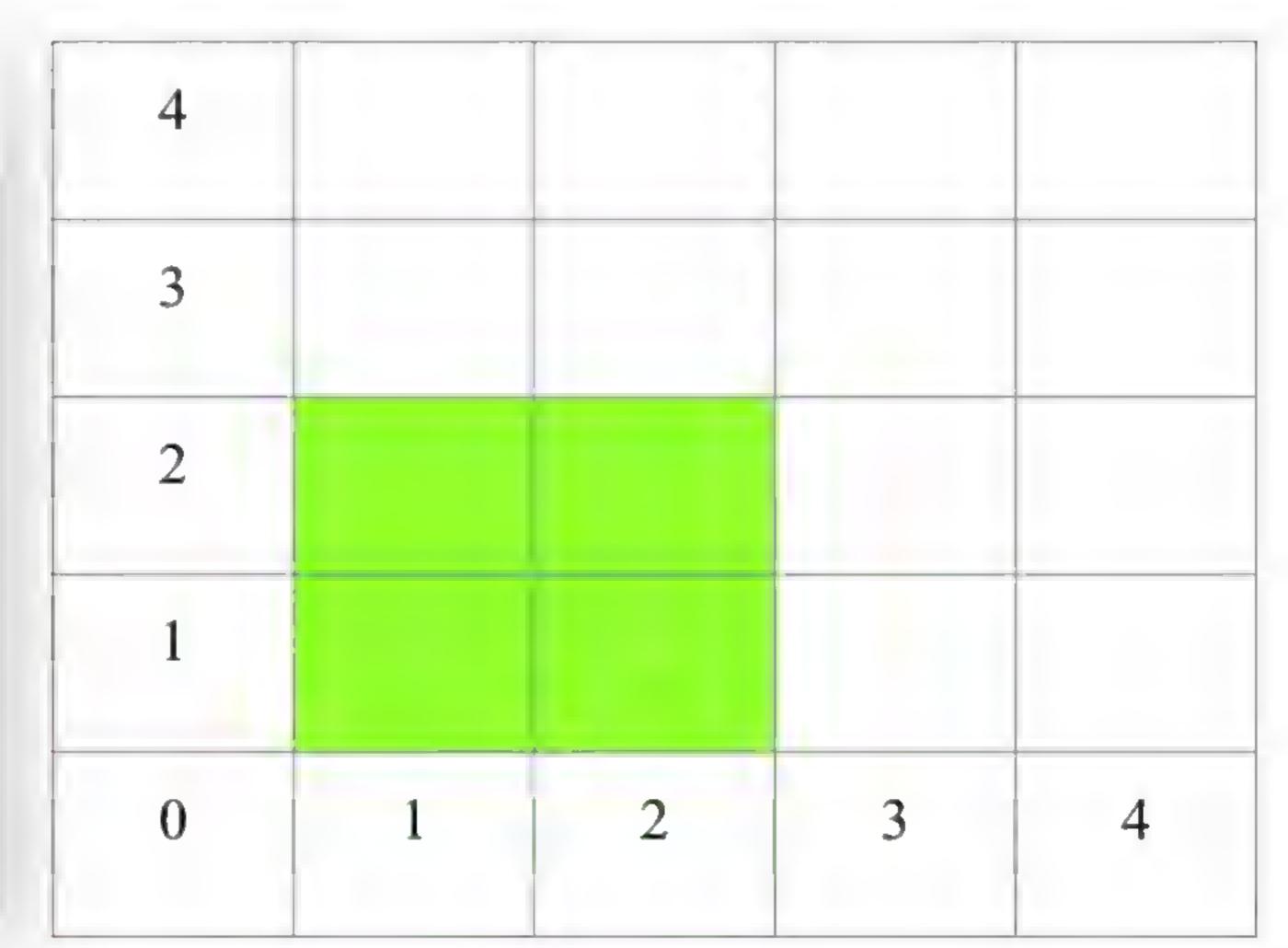






Как сейчас?

```
1 space:create_index('primary', {
2     type = 'tree',
3     parts = {{1, 'unsigned'}}
4 })
5 for i = 0, 4 do
6     for j = 0, 4 do
7         space:insert({i * 5 + j, i, j})
8     end
9 end
```



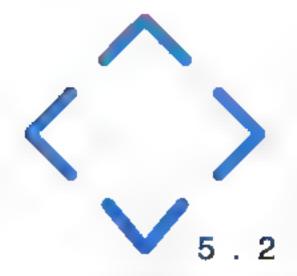




Наивный подход

```
1 result = {}
2 for _, tuple in space:pairs() do
3    local x = tuple[2]
4    local y = tuple[3]
5    if (1 <= x and x <= 2) and (1 <= y and y <= 2) then
6        table.insert(result, tuple)
7    end
8 end</pre>
```





B-Tree

```
index = space:create_index('secondary', {
     type = 'tree',
     parts = {{2, 'unsigned'}, {3, 'unsigned'}}
  result :- {}
  Local H = tuple[2]
local y = tuple[3]
    1 m >= 3 mm y > 3 mm
        DEGOR
     ena
13
    table.insert(result, tuple)
     and
Ta end
```

B-Tree

```
indem = space:create_indem('secondary', {
       type = 'tree'
       parts = {{2, 'unsigned'}, {3, | | | | }}
 6 result = {}
7 for _, tuple in index:pairs({1, 1}, 'GE') do
 8
       local x = tuple[2]
       local y = tuple[3]
10
11
       if x \ge 2 and y \ge 2 then
           break
12
13
       end
14
15
       if 1 \le y and y \le 2 then
           table.insert(result, tuple)
16
       end
18 end
```

```
space:create_index('primary', {
      type = 'tree',
      parts = {{1, 'unsigned'}}
  for i = 0, 4 do
      for j = 0, 4 do
          space:insert({i * 5 + j, {i, j}})
       end
10 end
   index = space:create index('secondary', {
     type 'rtree',
      parts = {{2, 'array'}},
  dimension = 2,
})
result = index:select({1, 1, 2, 2}, TE)
```

```
space:create_index( property, {
     type = -- ,
     parts = {{1, unsigned}}}
i = 0, A
    j = 0, 4 da
        space:insert({i * 5 + j, {i, j}})
     end
  end
  index = space:create index(' , {
    type == rtree,
     parts = {{2, _____}}},
   dimension = 2,
})
```

```
space:create_index( property, {
      type = = ;
      parts = {{1, | | | | | }}
i = 0, 4 = 0
      j = 0, 4 dn
          space:insert({i * 5 + j, {i, j}})
12 index = space:create_index('secondary', {
      type = 'rtree',
13
      parts = {{2, 'array'}},
      dimension = 2,
15
16 })
  result = index:select({1, 1, 2, 2}, IE)
```

```
space:create_index( pirm', {
     type = tim',
     i = 0, A
     j = 0, 4 da
        space:insert({i * 5 + j, {i, j}})
    nind
  index = space:create_index(' , {
     type = refer
     parts = {{2, ____'}},
  dimension = 2,
})
18 result = index:select({1, 1, 2, 2}, 'LE')
```

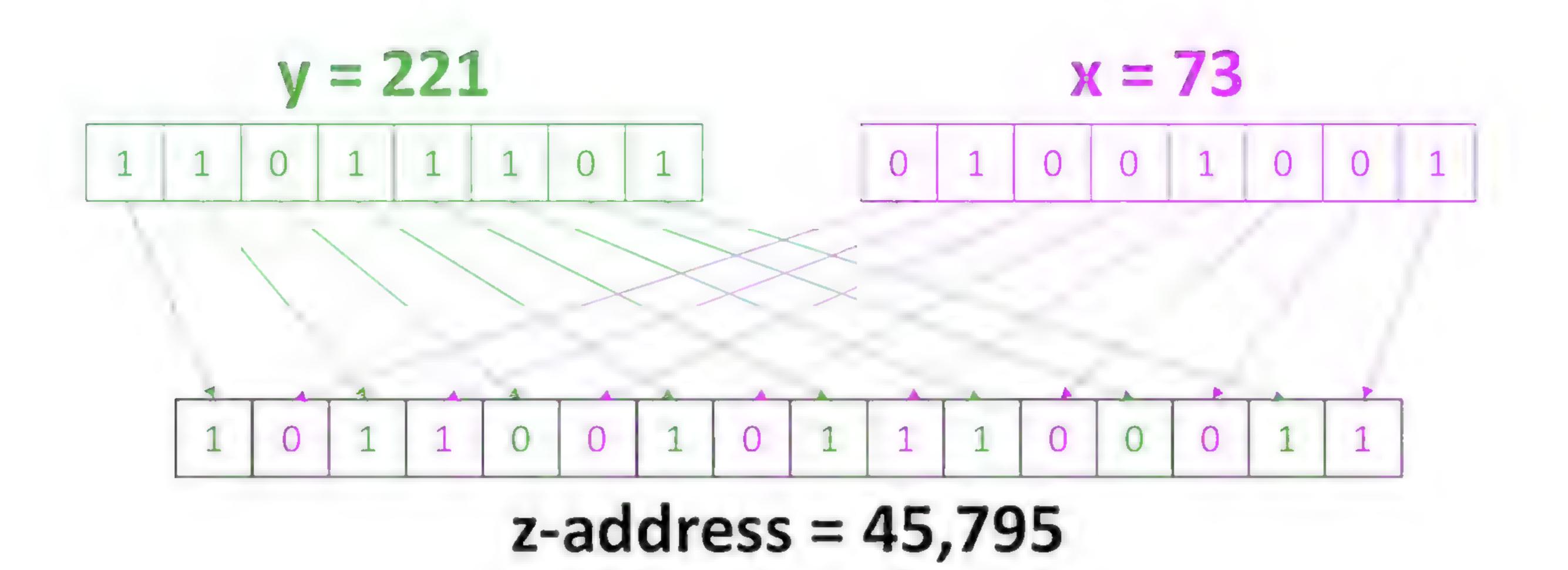
```
space:create_index('primary', {
       type = 'tree',
       parts = {{1, 'unsigned'}}
 6 	ext{ for } i = 0, 4 	ext{ do}
       for j = 0, 4 do
           space:insert({i * 5 + j, {i, j}})
       end
10 end
12 index = space:create_index('secondary', {
       type = 'rtree',
13
     parts = {{2, 'array'}},
15 dimension = 2,
16 })
18 result = index:select({1, 1, 2, 2}, 'LE')
```

Кривая Z-порядка

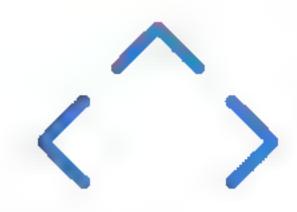
	X: 000	1 001	2 010	3 011	100	5 101	6 110	7 111	
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101	
1 001	000010	000011	000110	000111	010010	010011	010110	010111	
2 010	001000	001001	001100	001101	011000	011001	011100	011101	
3 011	001010	001011	001110	001111	011010	011011	011110	011111	
4 100	100000	100001	100100	100101	110000	110001	110100	110101	
5 101	100010	100011	100110	100111	110010	110011	110110	110111	
6 110	101000	101001	101100	101101	111000	111001	111100	111101	
7 111	101010	101011	101110	101111	111010	111011	111110	111111	







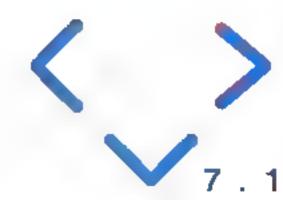




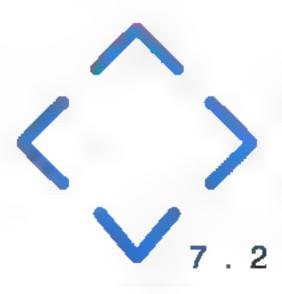
ПОИСК

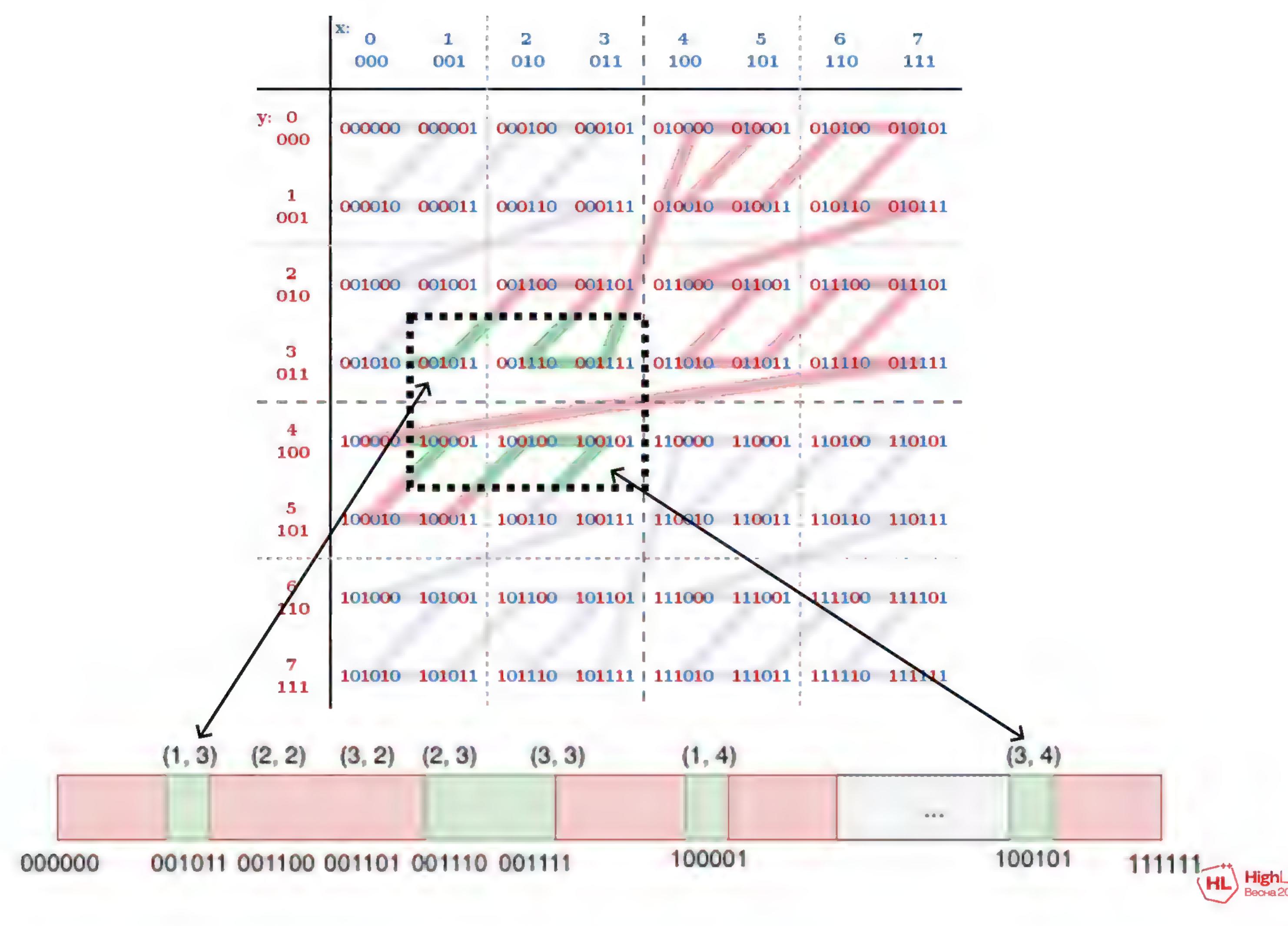
	x: 000	1 001	2 010	3 011	1 4 1 100	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101
1 001	000010	000011	000110	000111	010010	010011	010110	010111
2 010	001000	001001	001100	001101	011000	011001	011100	011101
3 011	001010	001011	001110	001111	011010	011011	011110	011111
4 100	100000	100001	100100	100101	110000	110001	110100	110101
5 101	100010	100011	100110	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	111000	111001	111100	111101
7 111	101010	101011	101110	101111	111010	111011	111110	111111

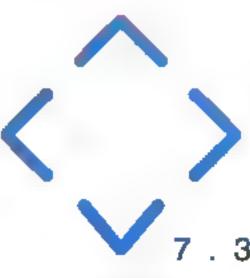


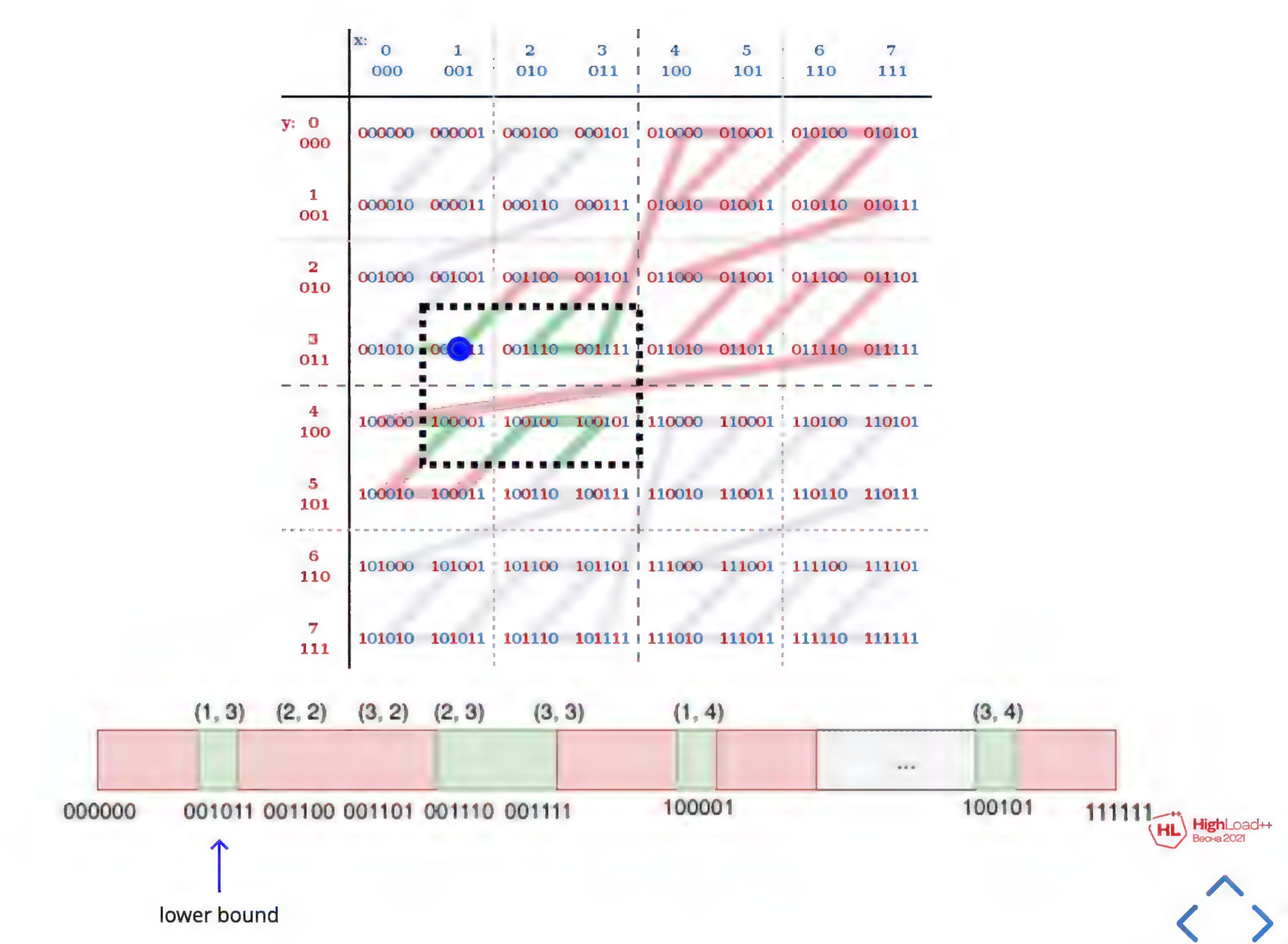


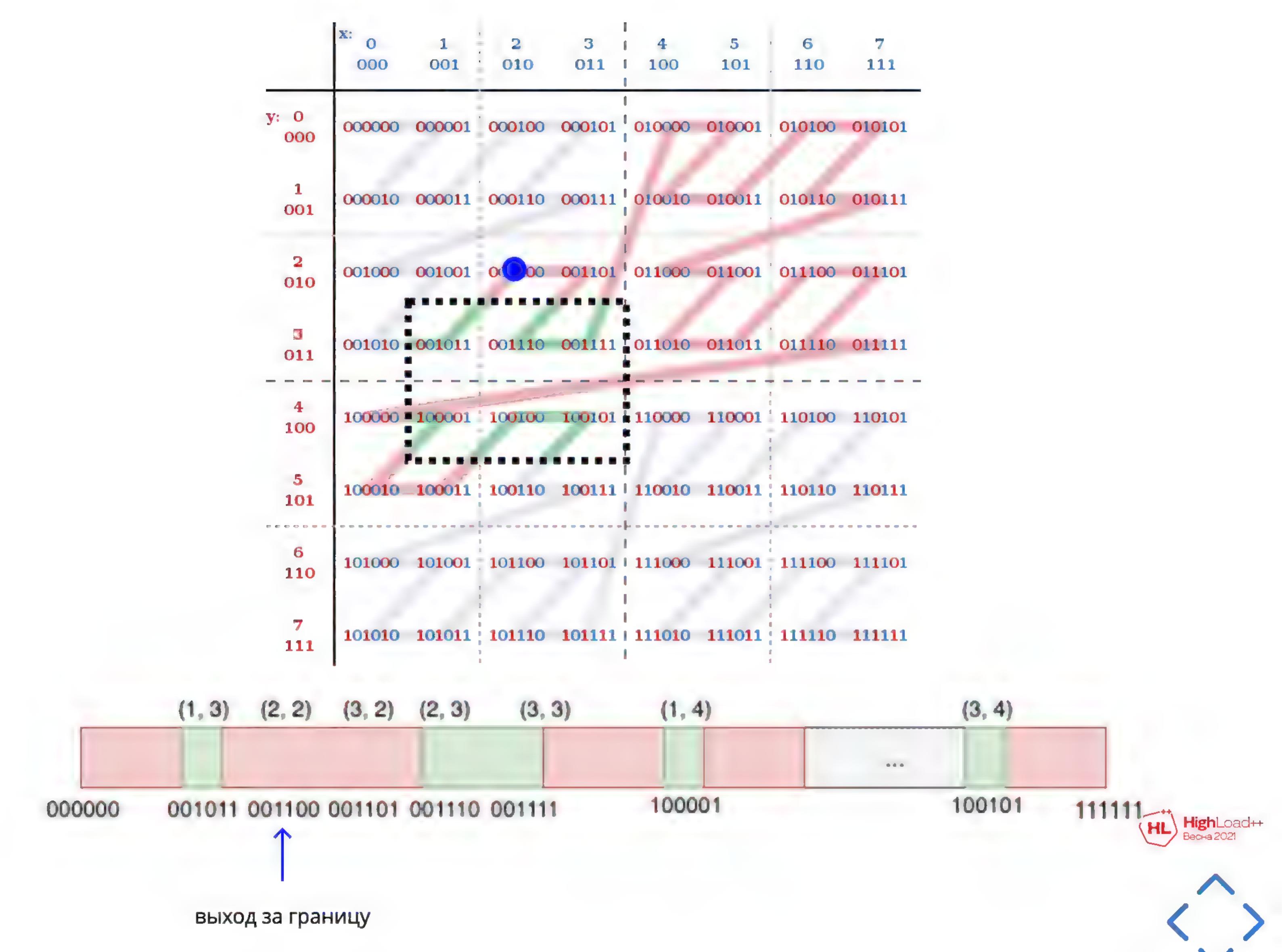
		000 000	1 001	2 010	3 011	1 4 1 100	5 101	6 110	7 111	
	y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101	
	1 001	000010			000111			010110	010111	
	2 010	001000	001001	001100	001101	011000	- Andrews	011100	011101	
	3 011	001010			001111		011011	011110	011111	
	4 100	100000	100001	100100	100101	110000	110001	110100	110101	
	5 101			100110	100111	1 110010	110011	110110	110111	
	6 110	101000	101001	101100	101101	1 111000	111001	111100	111101	
	7 111	101010	101011	101110	101111	111010	111011	111110	111111	
(1, 3)	(2, 2)	(3, 2)	(2, 3)	(3,	3)	(1,4	1)			(3, 4)

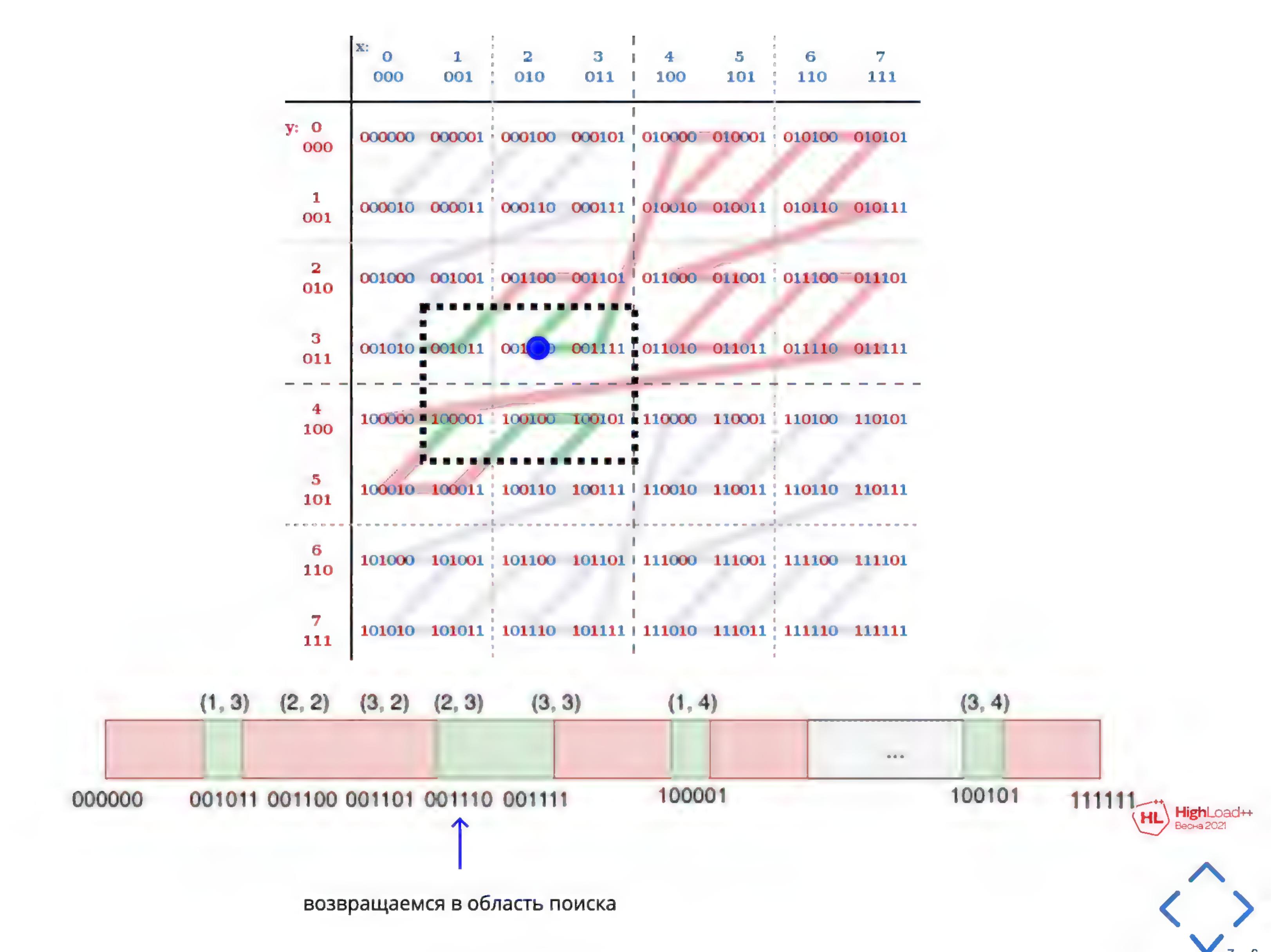


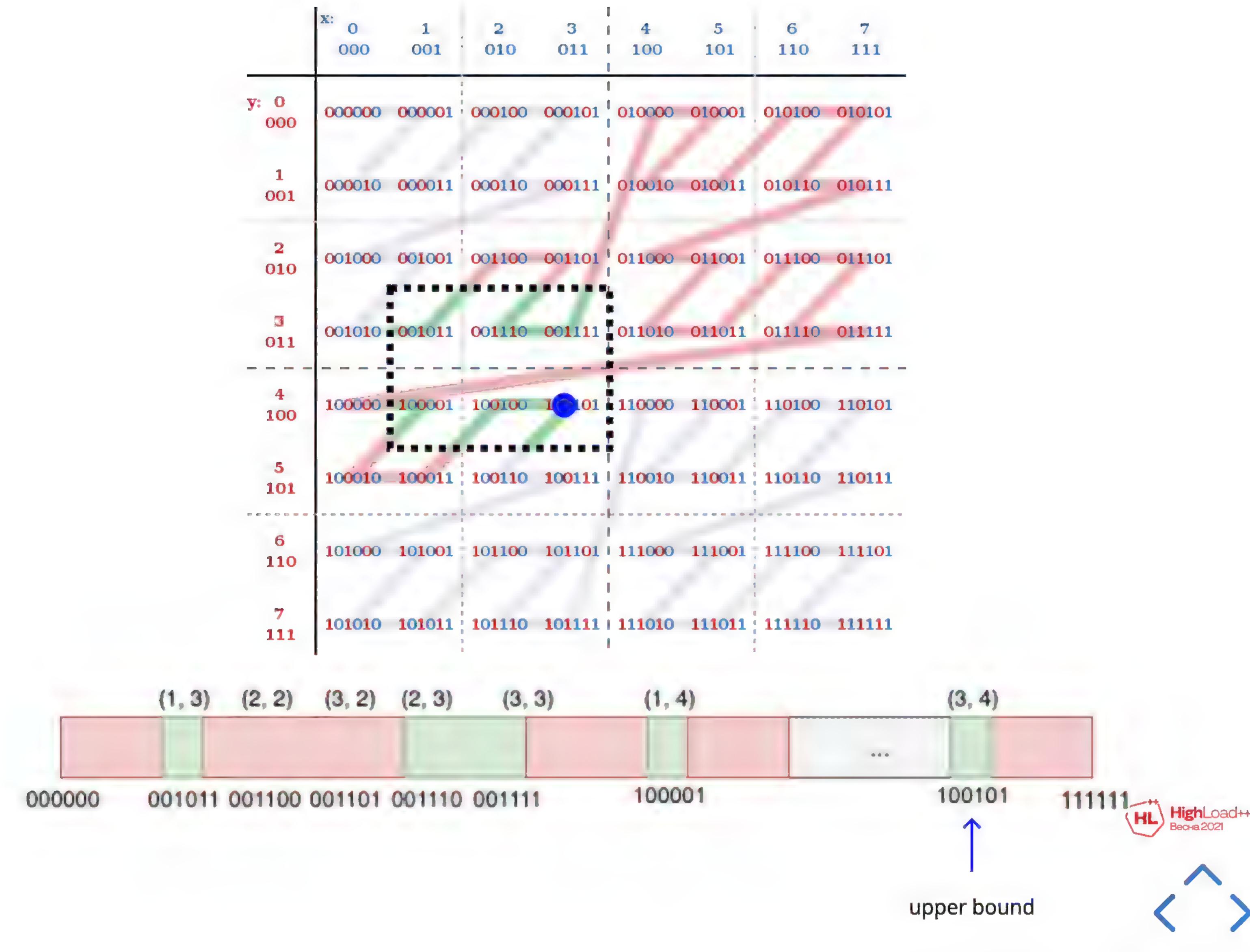












Встраивание в Tarantool

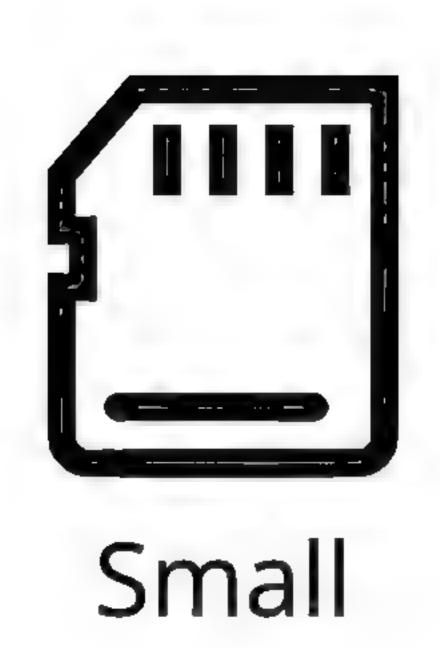
- Вычисление z-адреса
- Сохранение в В+*-Tree
- Реализация алгоритмов поиска

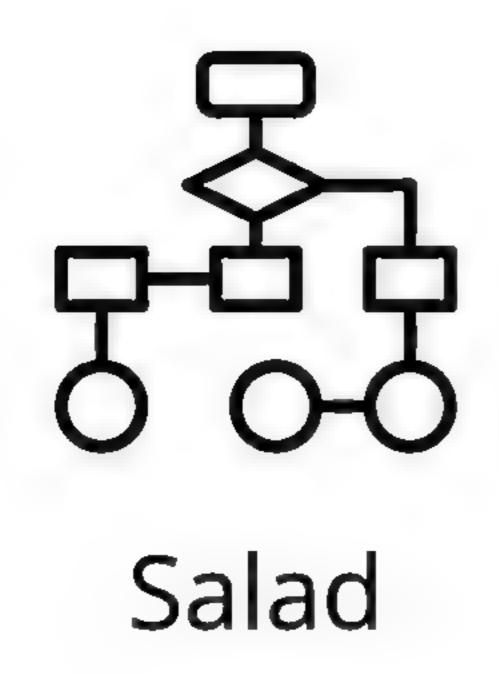




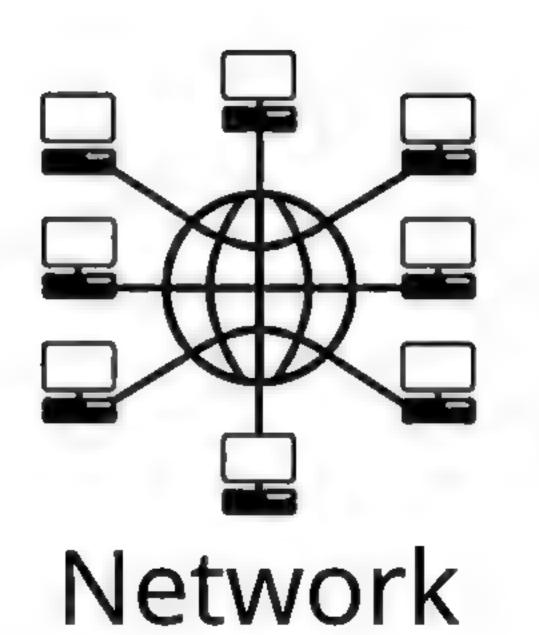


Что доступно разработчику?

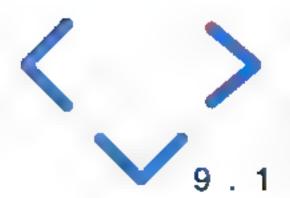








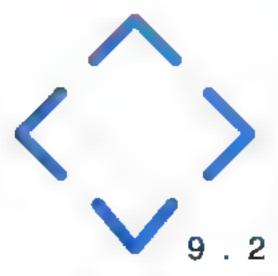




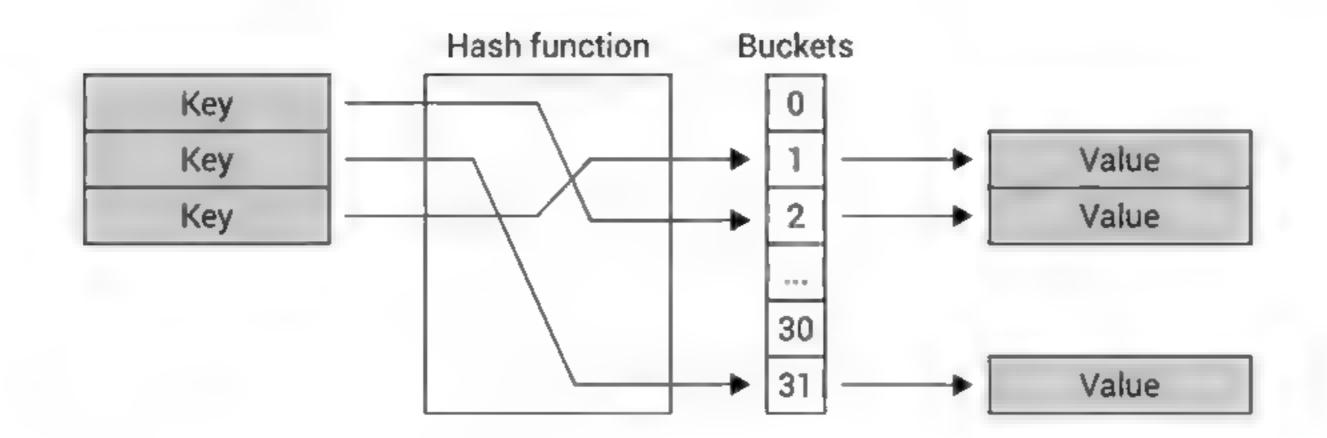
Specialized Memory ALLocators

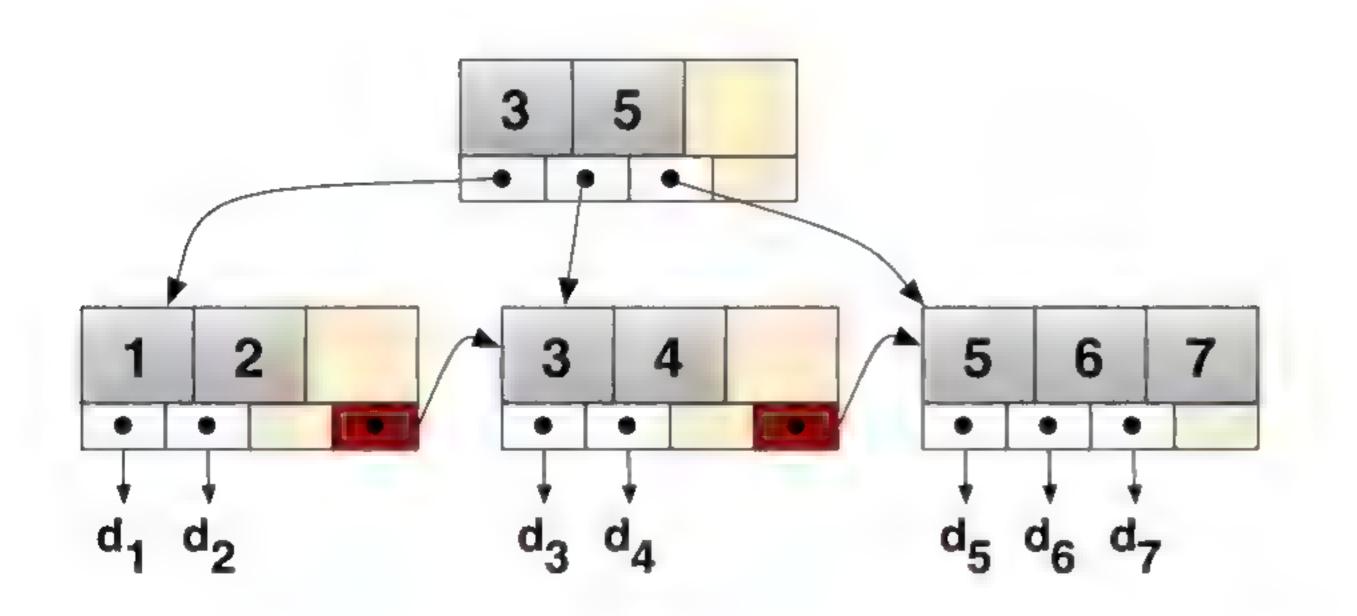


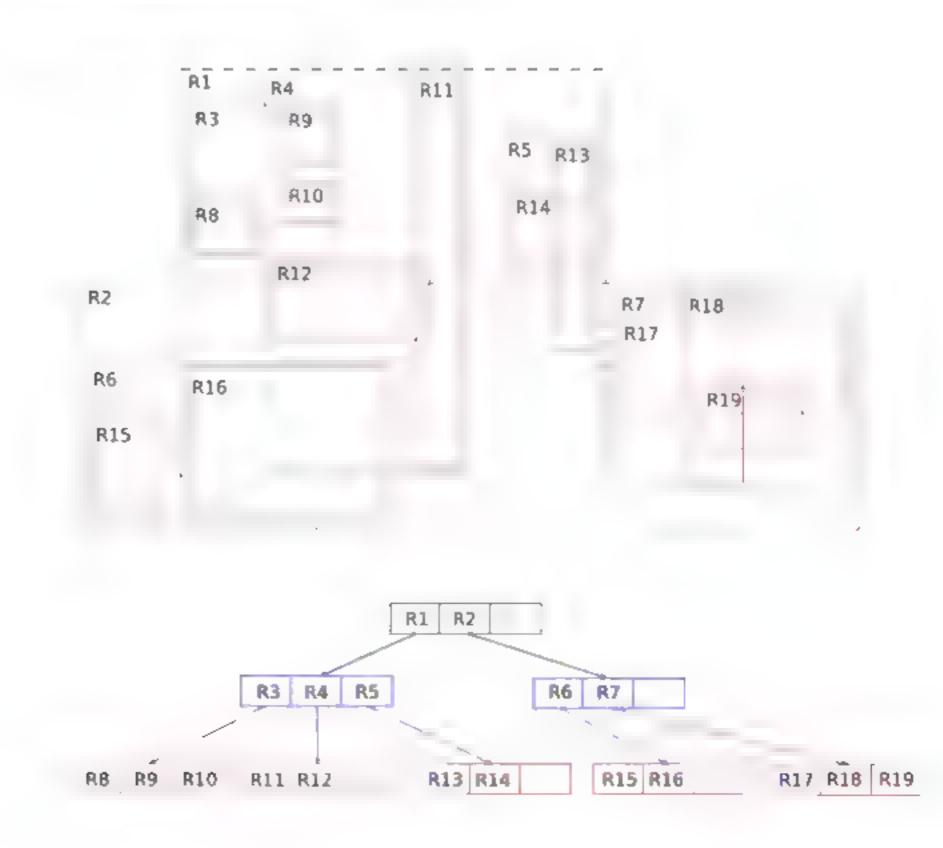




Some ALgorithms And Data structures











CAPI/Lua API

- Lua C API
- LuaJIT FFI
- Tarantool C API

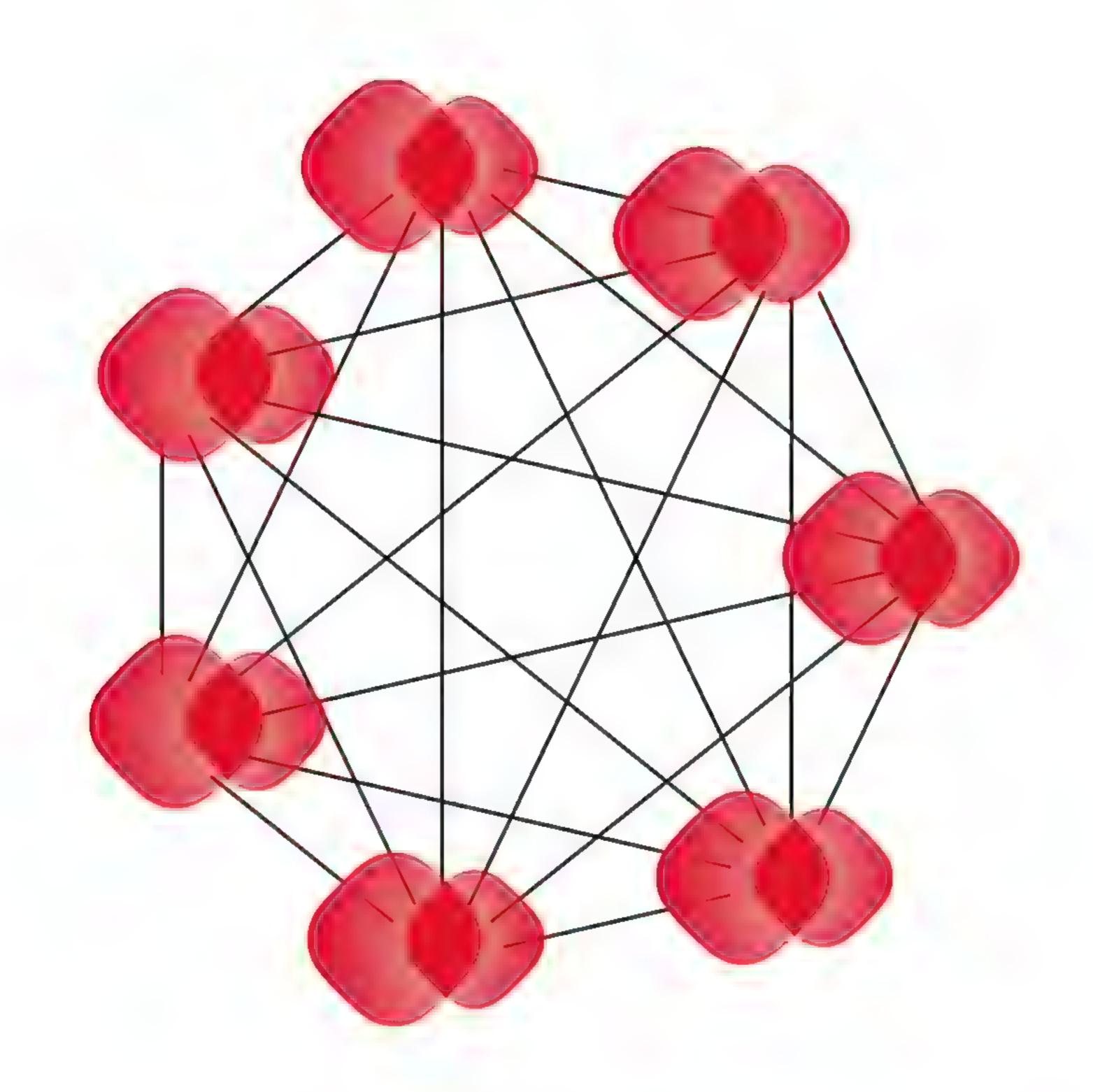




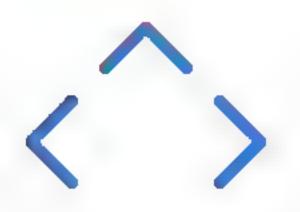


Network

- IProto
 - Call
 - Replication
 - DML
 - RAFT
- SWIM
- ...







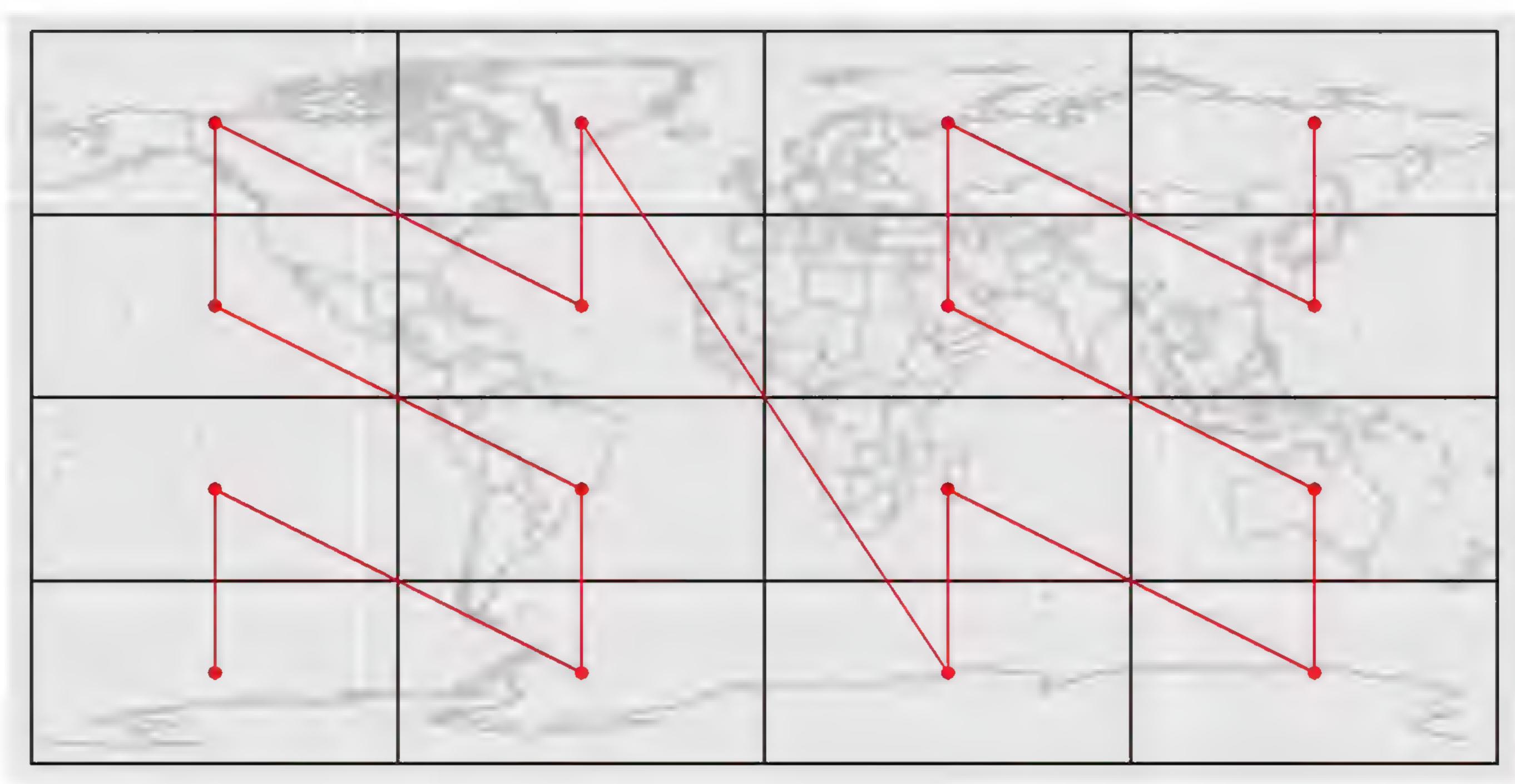
Что понадобилось?

- Набор алгоритмов для работы с Z-адресами
- B+*-Tree
- Bit array





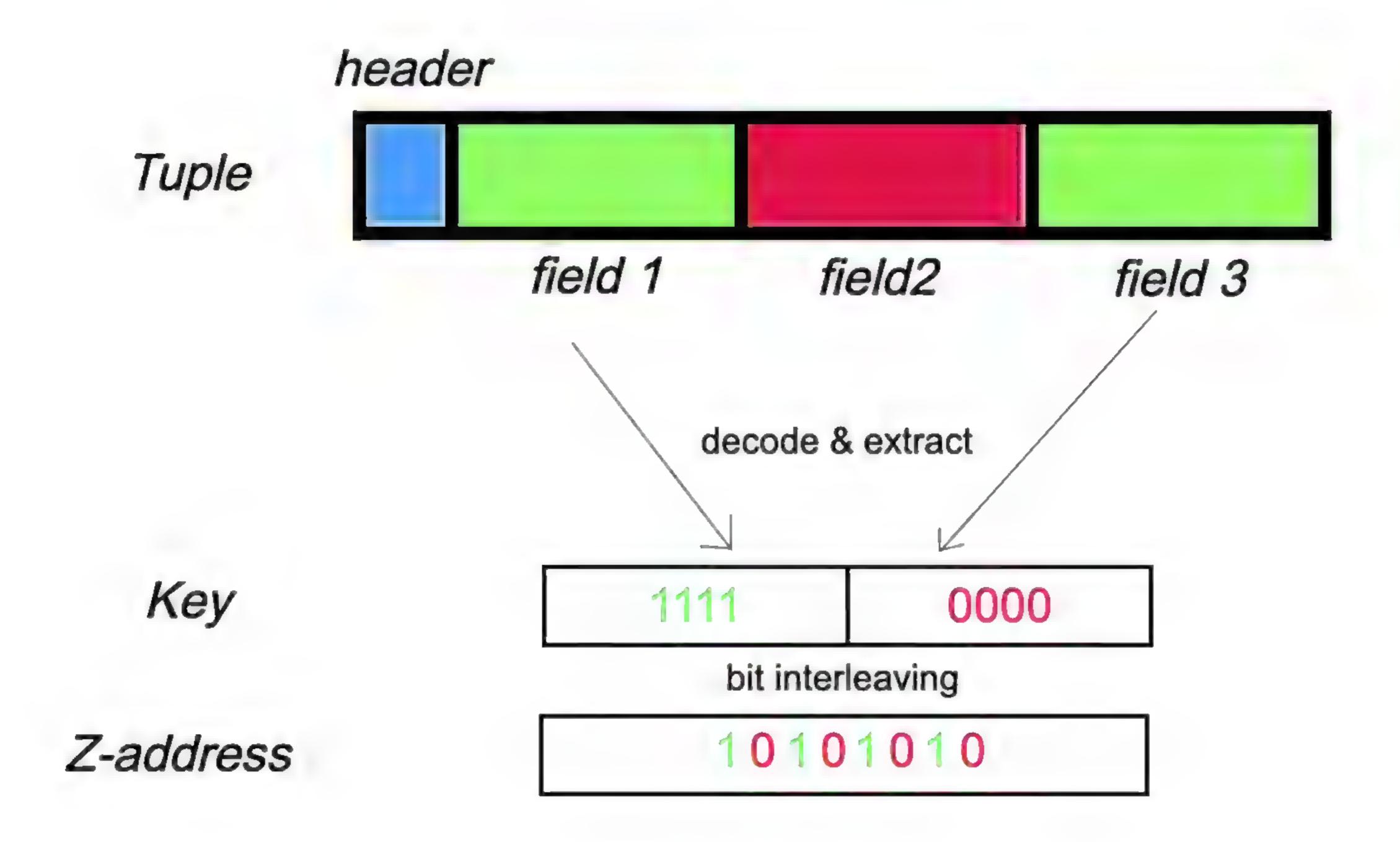
Работа с Z-адресами







Вычисление







Хранение в BPS-Tree

```
struct memtx_tree_data {
    /* Tuple that this node is represents. */
    struct tuple *tuple;
    /** Comparison hint, see key_hint(). */
    hint_t hint;
};
```

```
struct memtx_zcurve_data {
    /* Z-address. Read here: https://en.wikipedia.org/wiki/Z-order_curve */
    z_address *z_address;
    /** Tuple that this node is represents. */
    struct tuple *tuple;
};
```

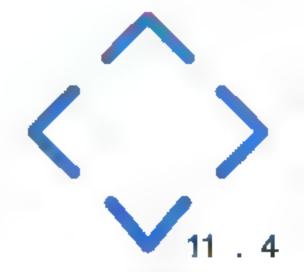




Работа с типами

Туре	Dec	Bin	Normalized	
unsigned	6	0000110	0000110	
integer	6	0000110	1000110	
	-6	10000110	0000110	
string	abc	0x61,0x62,0x63	0x61,0x62	
	a	0x61	0x61,0x00	





Next-jump-in

	x: n 000	1 001	2 010	3 011	1 100	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101
1 001	000010	000011	000110	000111	010010	010011	010110	010111
2 010	001000	001001	001100	001101	011000	011001	011100	011101
3 011	001010	001011	001110	001111	011010	011011	011110	011111
4 100	100000	100001	100100	100101	1110000	110001	110100	110101
5 101	100010	100011	100110	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	111000	111001	111100	111101
7 111	101010	101011	101110	101111	111010	111011	111110	111111

lower bound - 11 upper bound - 50

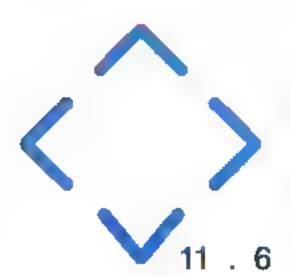




	x: 000	1 001	2 010	3 011	I I 4 I 100 I	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	1 010000	010001	010100	010101
1 001	000010	000011	000110	000111		010011	010110	010111
2 010	001000	001001	001100	001101	I		011100	011101
3 011	001010	001011	001110	001111	011010	011011	011110	011111
100	100000	100001	100100	100101	110000	110001	110100	110101
5 101	100010	100011	100110	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	1 1 111000	111001	111100	111101
7 111	101010	101011	101110	101111	1 1 111010	111011	111110	111111

lower bound - 11 upper bound - 50





	x: 0 000	001	2 010	3 011	1 4 1 100	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101
1 001	000010	000011	000110	000111	010010	010011	010110	010111
2 010	001000	001001	001100	001101	011000	011001	011100	011101
3 011	001010	001011	001110	001111	011010	011011	011110	011111
100	100000	100001	100100	100101	110000	110001	110100	110101
5 101	100010	100011	100110	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	1111000	111001	111100	111101
7 111	101010	101011	101110	101111	1 1 111010	111011	111110	111111

lower bound - 11 upper bound - 50 current - 40

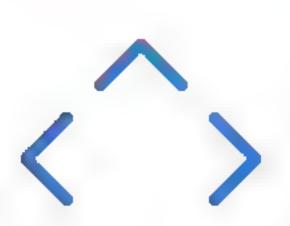




	x: 000	1 001	2 010	3 011	1 4 100 I	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101
1 001	000010		000110		010010			010111
2 010					011000			
3 011	001010	001011	001110	001111	011010	011011	011110	011111
4 100	100000	100001	100100	100101	110000	110001	110100	110101
5 101	100010	100011	1001.5	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	111000	111001	111100	111101
7 111	101010	101011	101110	101111	111010	111011	111110	111111

lower bound - 11 upper bound - 50 current - 40 NIP - 48

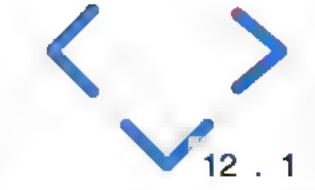




Индекс в Tarantool

- replace
- create_iterator
- get
- count
- size
- bsize
- min
- max
- update_def
- depends_on_pk
- •

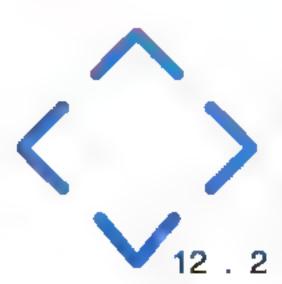




Replace

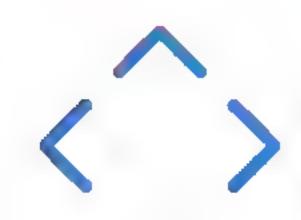
```
int
memtx_zcurve_index_replace(struct index *base, struct tuple *old_tuple,
struct tuple *new_tuple, enum dup_replace_mode mode,
struct tuple **result)
```





Create iterator

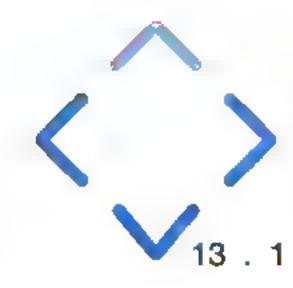




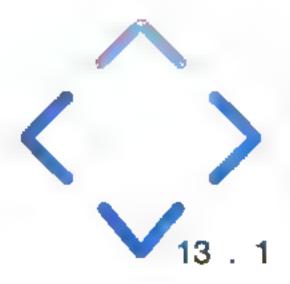




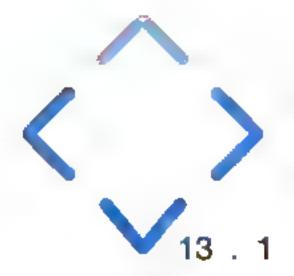






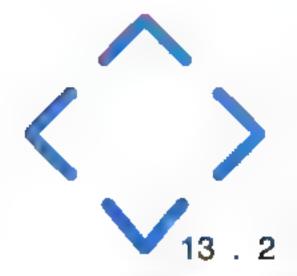






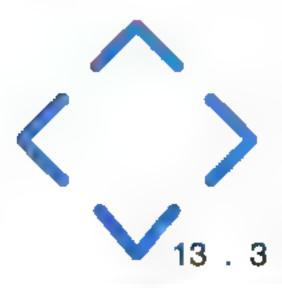
```
1 -- (2 <= x <= 3) and (3 <= y <= 5)
2 tarantool> secondary:select({2, 3, 3, 5})
3 ---
4 -- [15, 2, 3]
5 - [21, 3, 3]
6 - [16, 2, 4]
7 - [22, 3, 4]
8 - [17, 2, 5]
9 - [23, 3, 5]
```





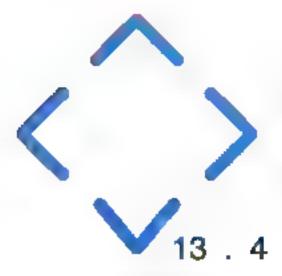
```
1 -- (x == 2) and (y == 3)
2 tarantool> secondary:select({2, 3})
3 ---
4 - - [15, 2, 3]
5 ...
```





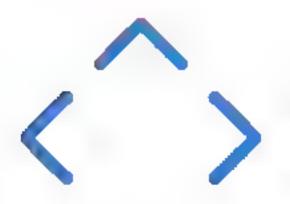
```
1 -- (2 \le x \le 3)
 tarantool> secondary:select({2, 3, box.NULL, box.NULL})
 - - [12, 2, 0]
  - [18, 3, 0]
  - [13, 2, 1]
   - [19, 3, 1]
    - [14, 2, 2]
    - [20, 3, 2]
    - [15, 2, 3]
    - [21, 3, 3]
    - [16, 2, 4]
    - [22, 3, 4]
    - [17, 2, 5]
    - [23, 3, 5]
```





```
1 -- (x >= 2) and (y >= 3)
 tarantool> secondary:select({2, box.NULL, 3, box.NULL})
 --[15, 2, 3]
  - [21, 3, 3]
  - [27, 4, 3]
    - [16, 2, 4]
    - [22, 3, 4]
    - [23, 3, 5]
    - [28, 4, 4]
    -[34, 5, 4]
    -[29, 4, 5]
    - [35, 5, 5]
```

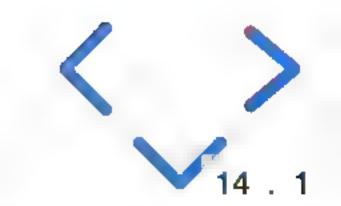




Z-order curve vs B-Tree

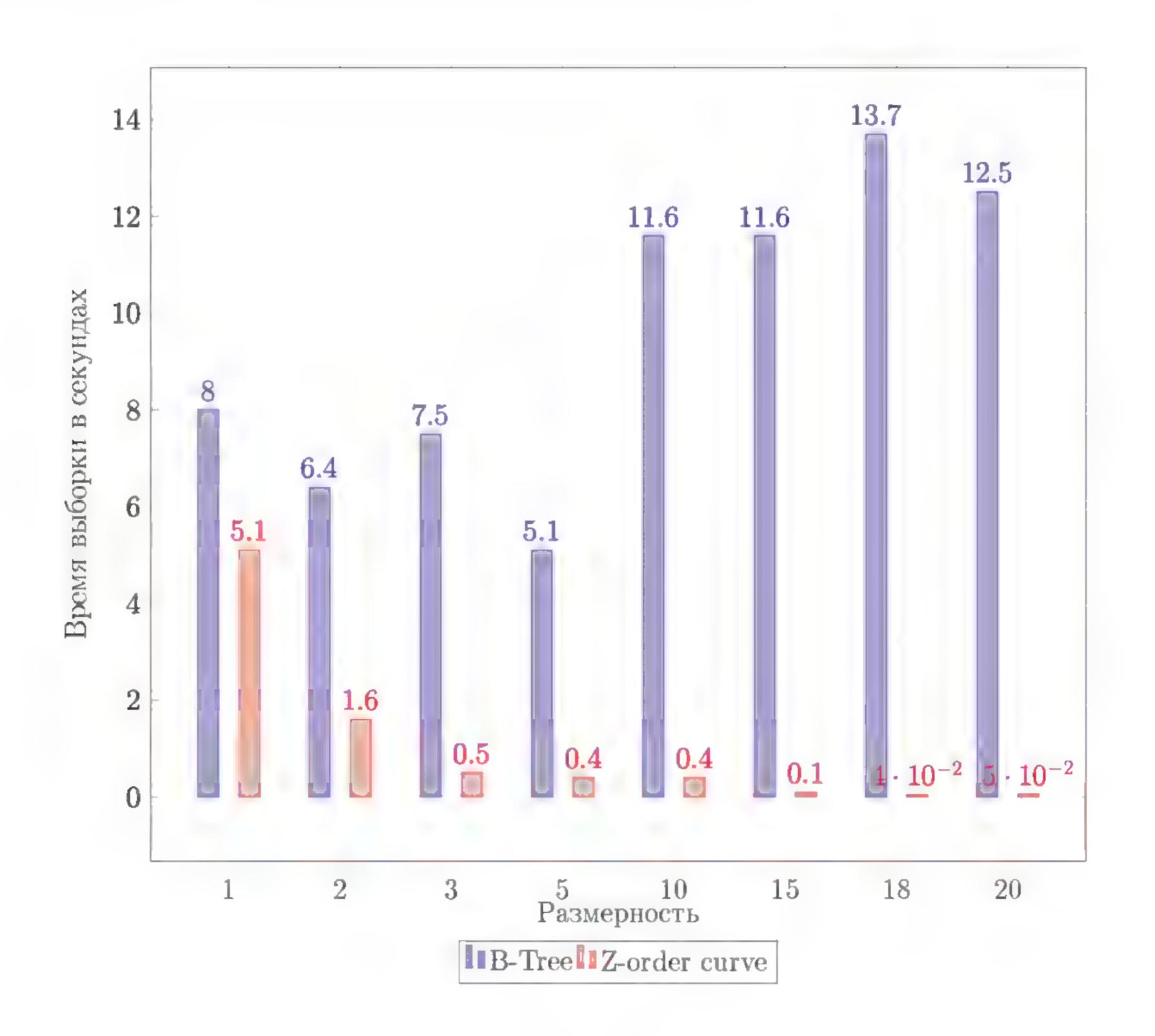
- Поиск в прямоугольной области
- Префиксный поиск



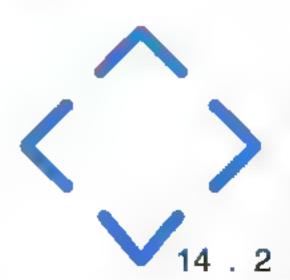


Поиск в прямоугольной области

10^6 точек, распределенных равномерно

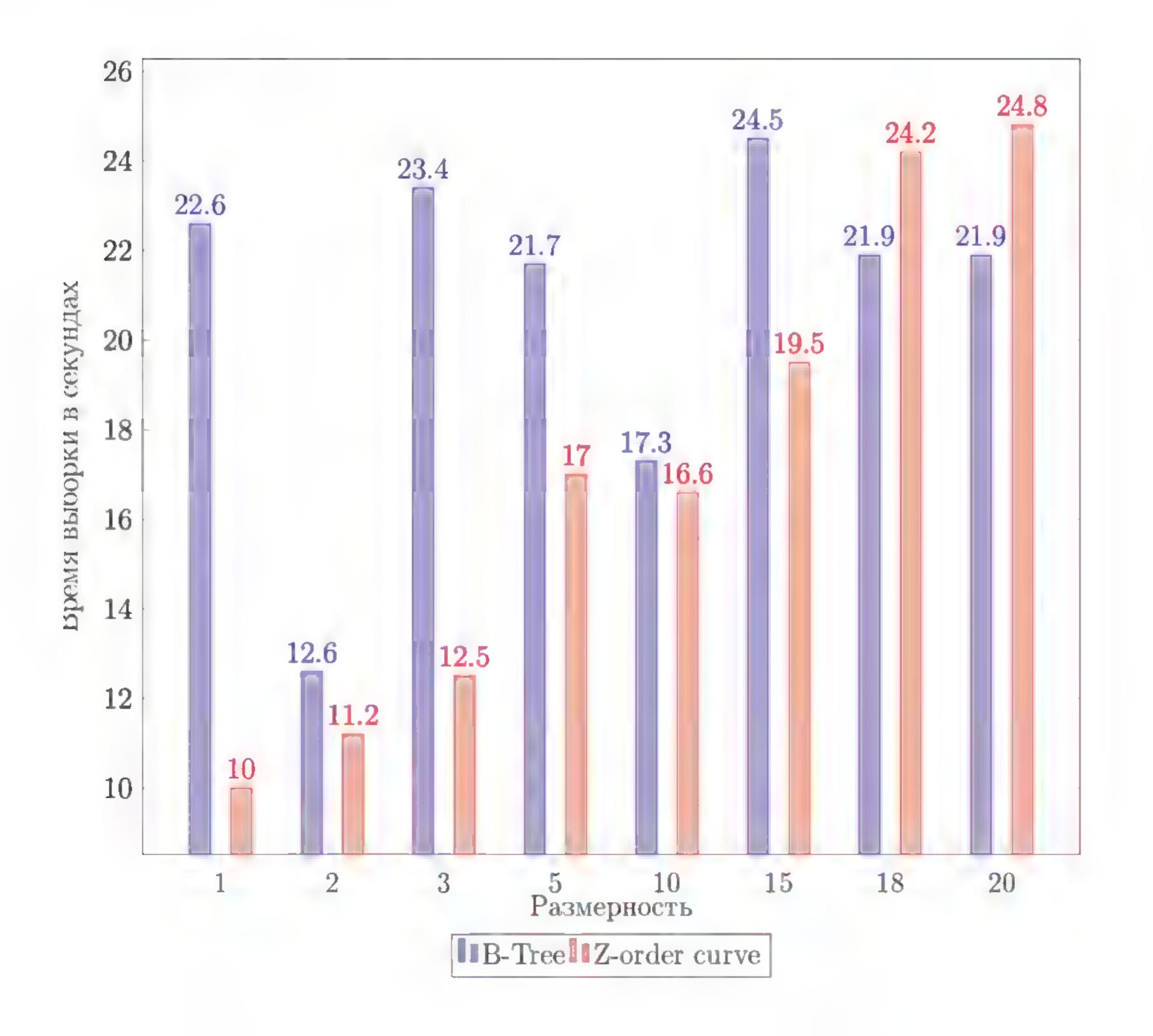




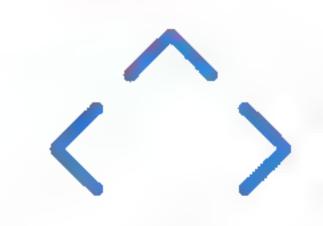


Поиск в прямоугольной области

10^6 точек, смещенных к 0



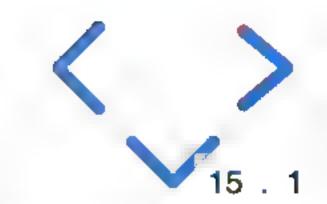




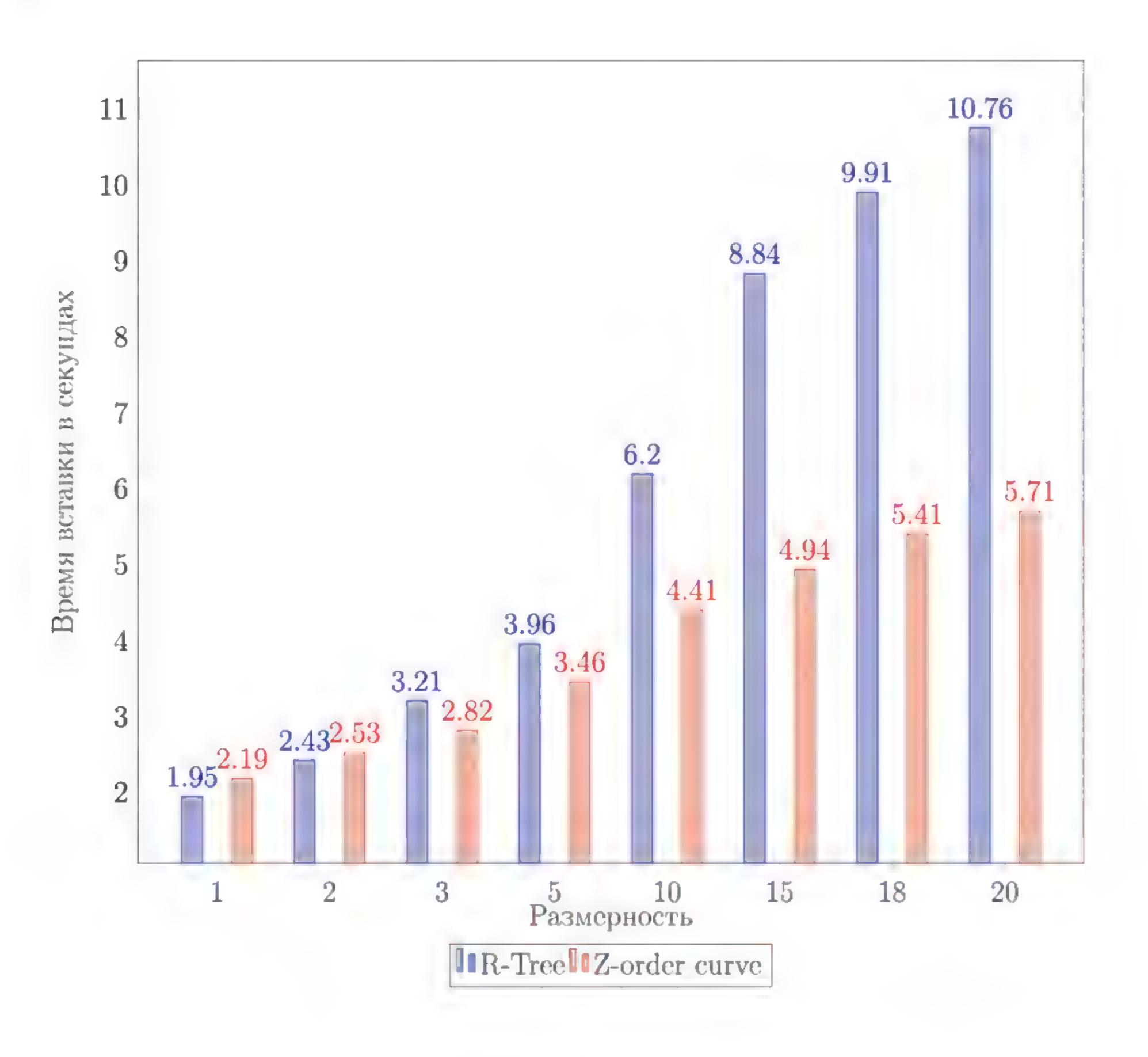
Z-order curve vs R-Tree

- Запись
- Потребляемая память
- Поиск в прямоугольной области

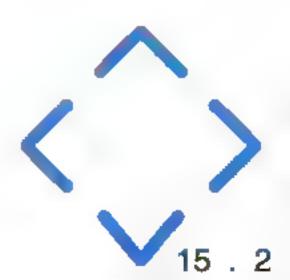




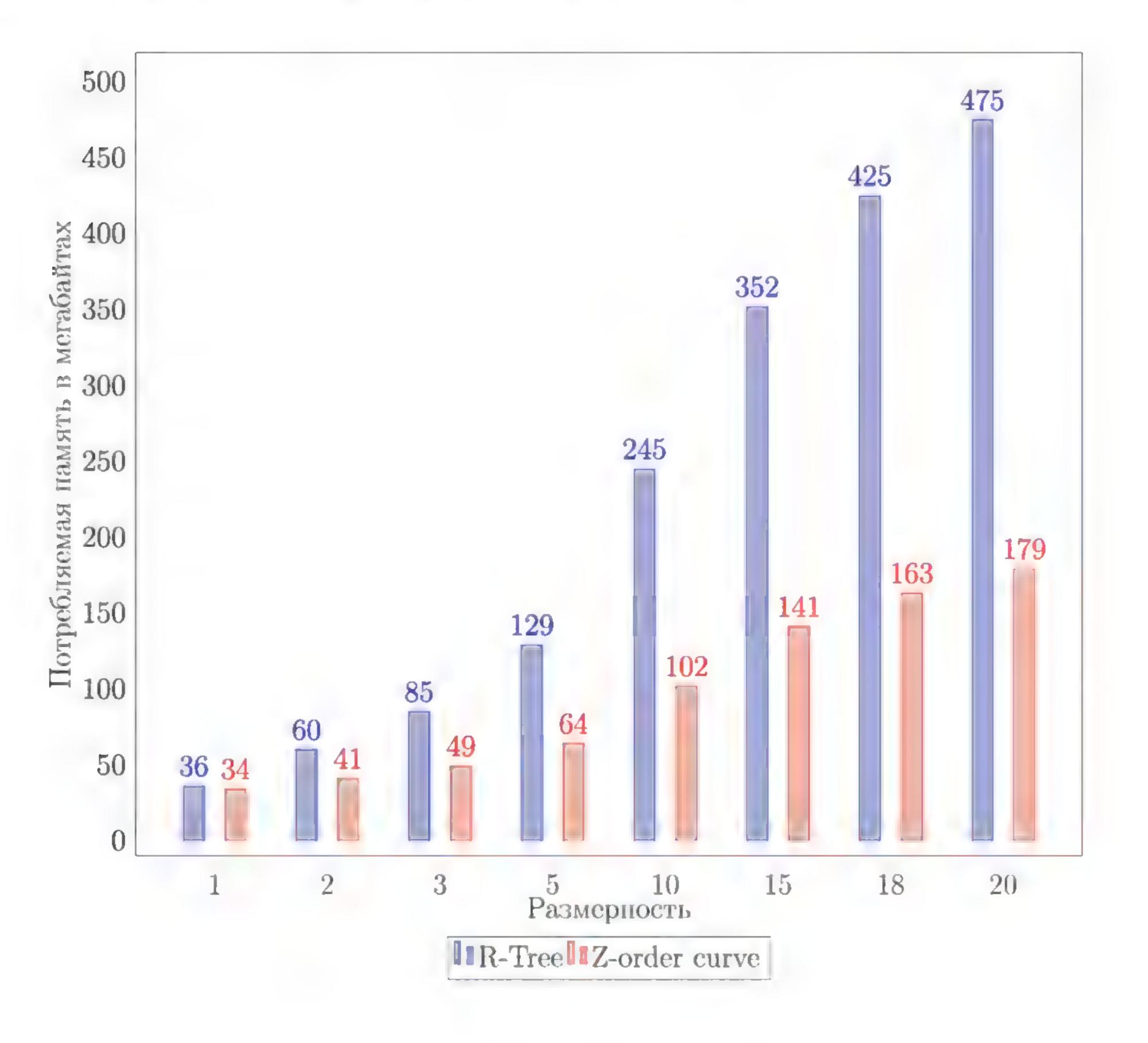
Запись



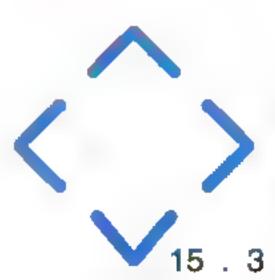




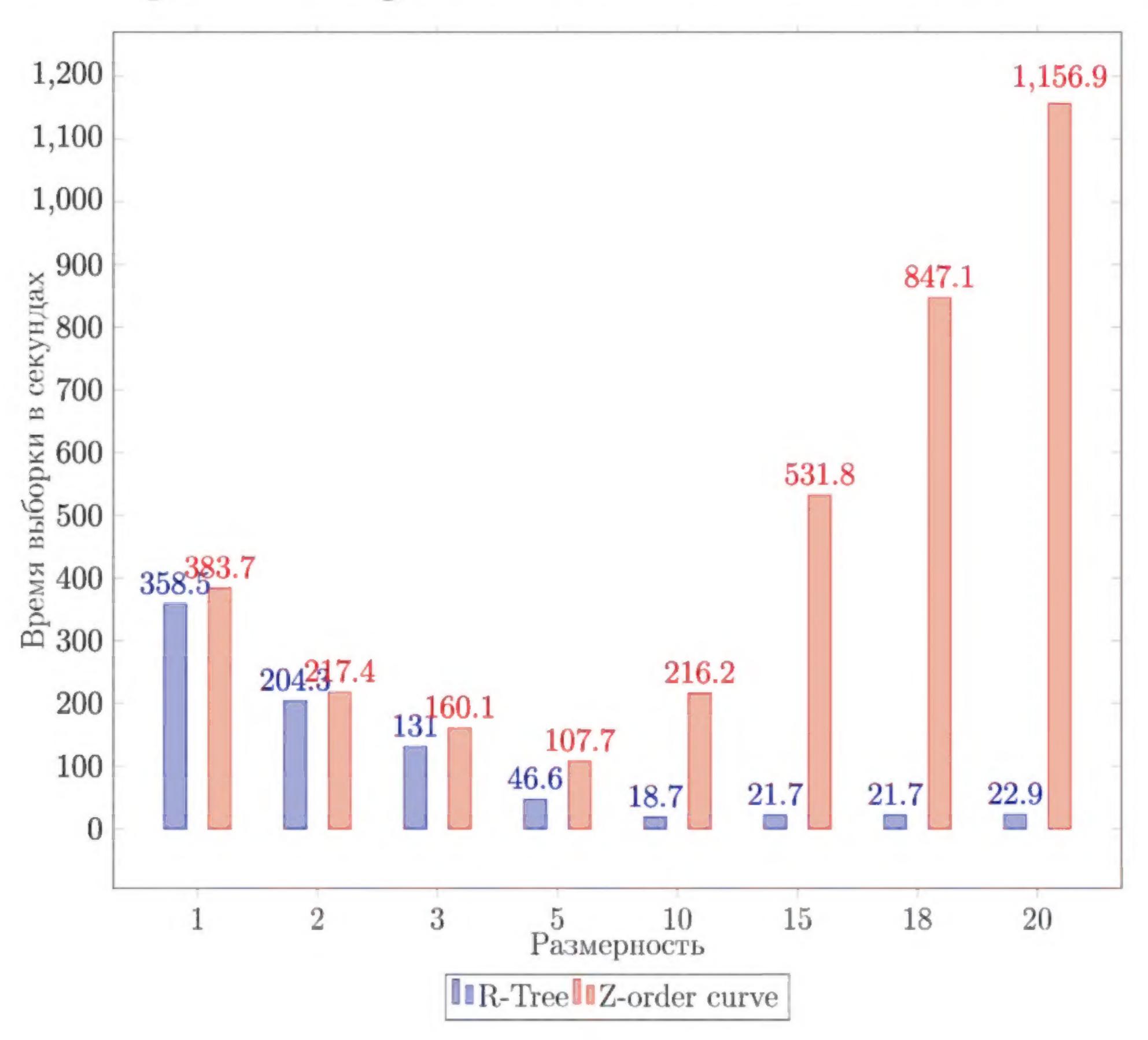
Потребление памяти



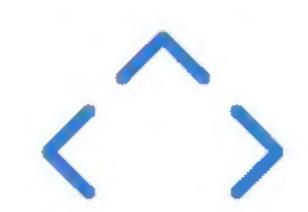




Поиск в прямоугольной области







Почему за год всё сломалось?



alexiyapunov 8 февраля 2021 в 16:45

Менеджер транзакций для базы данных в оперативнои памяти

Блог компании Mail.ru Group, Высокая производительность, Алгоритмы, Хранение данных, Tarantool







Вывод

- Отрицательный результат тоже результат :)
- Tarantool отличная платформа для изучения БД
- Не бойтесь экспериментировать





Вопросы

Олег Бабин

tg: @olegrok

e-mail: o.babin@corp.mail.ru

Fork:

https://github.com/olegrok/tarantool/tree/z-order-curve-index



